UNIVERSIDADE FEDERAL DO PARANÁ

GRAZIELLA ARAUJO DE OLIVEIRA LAPKOSKI

AVALIAÇÃO DE LEITURA EM INGLÊS PARA FINS ACADÊMICOS: ELABORAÇÃO E ANÁLISE DE TESTE DE SUFICIÊNCIA

CURITIBA

2008

GRAZIELLA ARAUJO DE OLIVEIRA LAPKOSKI

AVALIAÇÃO DE LEITURA EM INGLÊS PARA FINS ACADÊMICOS: ELABORAÇÃO E ANÁLISE DE TESTE DE SUFICIÊNCIA

Dissertação de mestrado, apresentado como requisito parcial à obtenção do grau de Mestre em Estudos Lingüísticos, do Curso de Pós-Graduação em Letras, do Setor de Ciências Humanas, Letras e Artes da Universidade Federal do Paraná.

Orientador: Prof. Dr. José Erasmo Gruginski

CURITIBA

2008

Ao meu marido, Geraldo, e minha filha, Thaís, pela ajuda, confiança e paciência; por terem aceito as longas e intermináveis horas de trabalho; pela compreensão nos inúmeros momentos em que, “abduzida” mentalmente pela leitura e escrita, não participei ativamente de suas vidas e pelo amor incondicional!

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. José Erasmo Gruginski, por ter conduzido com segurança todo o trabalho de orientação, pelos inúmeros ensinamentos, pela paciência, otimismo, incentivo e pela confiança em mim depositada!

À Profa. Dra. Vera Lúcia Posnik Rolloff, por ter dirigido minha atenção e incentivado a estudar a avaliação de leitura em língua estrangeira, por ter sido fonte de inspiração e exemplo de vida e pelas sábias observações e conselhos durante a qualificação e sempre que eu precisei.

Ao Prof. Dr. Michael Watkins, por ter contribuído para a conclusão desse trabalho com suas críticas e sugestões durante a qualificação!

A todos os professores que, gentilmente, disponibilizaram suas aulas para a aplicação dos meus testes com seus alunos. A todos os alunos que aceitaram participar da pesquisa, pelo simples desejo de ajudar.

Aos colegas do CELIN que voluntariamente utilizaram suas horas de folga para participar da pesquisa e responder ao teste, manifestando seu apoio!

Ao Odair, pela atenção e profissionalismo!

À Leonilda Procailo, Juliana Martinez e Denise Hibarino, por terem mantido suas mãos amigas estendidas durante todo o mestrado. Pelo encorajamento nos momentos de cansaço, pelo incentivo constante e pela ajuda inestimável na correção dos testes!

À minha mãe, Mazir, que me ensinou o valor do conhecimento e do estudo e me ajudou sempre que precisei!

Ao meu pai, José Carlos (in memoriam), pela forma de demonstrar apreciação e de valorizar cada uma das minhas pequenas e grandes conquistas!

RESUMO

Com base nas normas institucionais que prevêem a realização de testes de suficiência em língua estrangeira para os cursos de mestrado e doutorado e na necessidade de conferir maior confiabilidade e validade a essas avaliações, o presente estudo visa a elaboração, correção e análise de testes de suficiência de leitura em inglês para mestrado e doutorado. Ele procurou verificar a possibilidade de garantir, de forma mais clara aos alunos e à instituição, que o processo de avaliação é justo e válido. O estudo foi realizado com a elaboração de quatro testes-piloto com questões discursivas, que foram respondidas por grupos diferentes de indivíduos. A partir dos resultados observados, obtidos por meio do processo de correção e da analise Rasch, um dos testes-piloto foi, então, selecionado e teve algumas de suas questões reelaboradas, resultando em um teste final, que foi aplicado em um grupo diferente dos anteriores. Posteriormente, foi realizada uma análise contrastiva entre o teste selecionado e sua versão final, para verificar a eficiência do uso do Modelo Rasch para melhorar a qualidade do teste. Para minimizar os problemas inerentes à correção de questões discursivas foram estabelecidos alguns critérios para o processo de correção, como a elaboração de descritores em uma escala ordinal de classificação e o treinamento de corretores. Os testes-piloto foram corrigidos por dois corretores, sendo que cada um dos testes foi corrigido duas vezes, uma por cada corretor. O teste final foi corrigido por quatro corretores, sendo que cada teste foi também corrigido duas vezes, por dois corretores diferentes. A elaboração e a correção desses testes foram baseadas, especialmente, em Alderson (1995 e 2000), Bachman (1990), Bachman e Palmer (1996) e McNamara (1996 e 2000). Após a obtenção dos escores com base na escala de classificação, de acordo com Pasquali (2003) e Bond e Fox (2001 e 2007), foi utilizado o Modelo Rasch para converter a escala ordinal em escala intervalar e, a seguir, os dados e gráficos gerados foram analisados. Os resultados mostram que o estabelecimento de descritores, o treinamento dos corretores e a análise probabilística dos dados podem proporcionar um aumento na qualidade do processo de avaliação e conferir maior confiabilidade e validade aos testes.

Palavras-chave: compreensão de leitura; elaboração de teste de suficiência; análise de dados com o Modelo Rasch; processo de correção; confiabilidade; validade.

ABSTRACT

Based on the institutional rules which foresee carrying out foreign language sufficiency tests for Masters' and Doctorate degree courses and the need of giving those evaluations more reliability and validity, the present study aims at building, rating and analyzing reading sufficiency tests in English for academic purposes, particularly for master's and doctorate degrees. It sought to verify the possibility of ensuring that the evaluation process is reliable and valid, to the students and the institution, in a clearer way. The study was carried out with the design of four pilot tests with open-ended questions, which were answered by different groups of people. From the observed results, obtained through the rating process and the Rasch analysis, one of the pilot tests was selected and had some of its questions rewritten, resulting in a final test, which was answered by a different group. Later, a contrastive analysis between the chosen test and its final version was made, in order to check the efficiency of using the Rasch Model to improve the test quality. To minimize the problems inherent to the rating process of open-ended questions, some rating process criteria were established, such as drawing up some descriptors in an ordinal scale and training raters. The pilot tests were rated by two raters, given that each test was rated twice, once per each rater. The final tests were rated by four raters, given that each test was also rated twice, by two different raters. These tests building and rating were based mainly on Alderson (1995 - 2000), Bachman (1990), Bachman and Palmer (1996) and McNamara (1996 - 2000). After the scores were obtained based on the rating scale, in accordance with Pasquali (2003), and Bond and Fox (2001- 2007), the Rasch Model was used to convert the ordinal scale into an interval scale and, after that, the generated data and graphs were analyzed. The results showed that the drawing up of descriptors, the training of raters and the data probabilistic analysis can provide an increased quality of the evaluation process, making the tests more reliable and valid.

Keywords: reading comprehension; sufficiency test building; data analysis with Rasch Model; rating process; reliability; validity.

SUMÁRIO

1 INTRODUÇÃO 7
1.1 OBJETIVOS DE PESQUISA 10
1.1.1 Objetivo Geral 11
1.1.2 Objetivo Específico 11
1.2 JUSTIFICATIVA 12
1.3 ORGANIZAÇÃO 16
2 REVISÃO BIBLIOGRÁFICA DE LEITURA E AVALIAÇÃO 17
2.1 O QUE É COMPREENSÃO DE LEITURA 19
2.2 LEITOR IDEAL / LEITOR REAL 29
2.3 TEXTO 35
2.3.1 O Conceito de Gênero 37
3 REVISÃO BIBLIOGRÁFICA DE AVALIAÇÃO 42
3.1 MEDIÇÃO – AVALIAÇÃO – TESTE 42
3.2 O PROCESSO DE AVALIAÇÃO 45
3.2.1 Validade 46
3.2.1.1 Adequação do Teste Quanto Aos Objetivos Institucionais 48
3.2.2 Validade de construto 53
3.2.3 Validade de Conteúdo 62
3.2.4 Validade de Critério 64
3.2.5 Confiabilidade 68
3.2.5.1 Adequação quanto à Medida 71
3.2.5.2 Necessidade de uma Escala de Medição 72
3.2.5.3 A Natureza das Medidas e as Escalas 76
3.3 A QUESTÃO DO ERRO NA MEDIÇÃO: COMO TRABALHAR 80
3.4 A TEORIA CLÁSSICA DE TESTES (TCT) 82
3.5 O ESCORE BRUTO 83
3.6 O ESCORE PADRONIZADO E PERCENTIL 86
3.7 A TEORIA DE RESPOSTA AO ITEM (TRI) E O MODELO RASCH 89
3.8 DIFICULDADE DO ITEM 95
3.9 CORREÇÃO DE TESTES 97
3.9.1 Variabilidade relacionada ao leitor 98
3.9.2 Variabilidade relacionada à tarefa (e ao texto) 102
3.9.3 Variabilidade relacionada ao corretor 110
3.9.4 Treinamento de Corretores 115
4 DESENVOLVIMENTO E APLICAÇÃO DOS TESTES 118
4.1 METODOLOGIA DE PESQUISA 119
4.2 DESENVOLVIMENTO DO TESTE 124
4.2.1 Definição do Construto 125
4.2.2 Critérios de Correção 127

4.2.3 Unidimensionalidade do teste 129
4.2.4 Escolha de Textos 129
4.2.4.1 Resumos de teses e dissertações 130
4.2.4.2 Textos Científicos 134
4.2.5 Elaboração das Questões 137
4.3 CORREÇÃO DOS TESTES-PILOTO 143
4.3.1 – Problemas Identificados na Elaboração das Questões 144
4.3.2 Problemas Observados nas Respostas 148
4.4 ESCOLHA, REELABORAÇÃO E APLICAÇÃO DO TESTE 152
4.4.1 Escolha do Teste e Re-elaboração do Teste 153
4.4.2 Aplicação do teste 158
5 CORREÇÃO DOS TESTES E ANÁLISE DOS RESULTADOS 160
5.1 PROCESSO DE CORREÇÃO DO TESTE FINAL 161
5.1.1 Escala de Classificação e Descritores 162
5.1.2 Treinamento de Corretores 168
5.1.3 Processo de Análise da Correção – O Modelo Rasch (TRI) 169
5.1.4 Curva Característica do Item (CCI) 171
5.1.5 A escala intervalar 175
5.2 ANÁLISE DOS RESULTADOS - PILOTO 1B X TESTE FINAL 176
5.2.1 Piloto 1B 179
5.2.2 Piloto 1B X Teste Final 188
6 CONSIDERAÇÕES FINAIS 199
6.1 LIMITAÇÕES DO ESTUDO 202
6.2 SUGESTÕES PARA FUTURAS PESQUISAS 203
REFERÊNCIAS BIBLIOGRÁFICAS 206
GLOSSÁRIO 210
APÊNDICES 217
ANEXOS 259

1 INTRODUÇÃO

A área de avaliações vem crescendo consideravelmente em todos os setores da sociedade, que buscam métodos científicos para embasar suas decisões em argumentos sustentáveis pela teoria, pela prática ou por ambas. Cada vez mais empresas, governamentais e não governamentais, solicitam algum tipo de teste, seja de inteligência, de algum tipo de conhecimento global, ou de conhecimentos específicos, para a admissão de funcionários nos mais variados cargos. Avaliações que medem a inteligência, as que medem alguma capacidade intelectual de realização de tarefas, que buscam reconhecer as variáveis que interferem no desempenho dessa realização, entre inúmeras outras vêm sendo pesquisadas já há algumas décadas. Provavelmente, porque as avaliações têm sido cada vez mais utilizadas como base para tomada de decisões fundamentais na vida dos indivíduos, para propósitos tais como os educacionais, os de seleção e de certificação. O grande número de pessoas submetidas a esses testes impõe cada vez mais a necessidade de tornar o processo qualitativamente melhor tanto em relação à sua validade quanto à sua confiabilidade.

As avaliações de linguagem, especificamente, tem sido o foco de atenção de vários autores, como Andrew Cohen (1994), Tim McNamara (1996 - 2000), Lyle Bachman (1990 – 1996) e Adrian Palmer (1996), entre tantos outros de igual relevância no mundo acadêmico internacional. Entre as quatro habilidades de linguagem: fala, escrita, audição e leitura, esta última é amplamente utilizada nos mais diferentes meios educacionais e profissionais. A leitura em uma língua estrangeira, daqui em diante referida apenas como LE, em particular o inglês atualmente, é fundamental para todo indivíduo que pretende ampliar seus conhecimentos e possibilidades, quer nos estudos quer na profissão, por proporcionar acesso a um expressivo número de informações disponíveis nos meios eletrônicos, nos mais diferentes idiomas, principalmente em inglês.

As avaliações de leitura como LE, dessa forma, têm sido intensamente estudadas por vários pesquisadores como Charles Alderson (1995 - 2000),

Christine Nuttall (2000), Rosa Maria Nery (2003) e Celso Tumolo (2005), para citar apenas alguns.

Kintsch e Kintsch, por exemplo, afirmam que

> os objetivos da compreensão de texto[1] variam amplamente. O motivo para ler um manual pode ser para aprender como realizar uma atividade; pode-se ler uma história de detetive para se distrair por um tempo; pode-se ler o jornal para ficar informado, e assim por diante. Mas, em contextos educacionais, o objetivo é freqüentemente aprender a partir de um texto, isto é, construir um modelo situacional[2] que será lembrado e poderá ser usado eficazmente quando a informação fornecida pelo texto for necessária de alguma maneira mais tarde[3]. (2004, p.76, tradução nossa)

Independente da língua em que é escrito, geralmente textos científicos são considerados difíceis pelos estudantes, devido à profusão de termos técnicos, os inúmeros componentes que são citados e as relações entre todos e dada um deles, os sistemas, símbolos, fórmulas, etc. Em se tratando de textos acadêmicos, escritos para relatos de pesquisas de campo ou estudos teóricos, por exemplo, a complexidade aumenta consideravelmente. De acordo com Graesser, Léon, Otero (2002, p. 02) "os problemas são especialmente importantes para leitores com pouco conhecimento científico. Na verdade, todas as dificuldades são exacerbadas pelo fato de que a maioria dos

---

[1] Como esta pesquisa trata especificamente de avaliação de leitura em inglês como língua estrangeira, o termo texto será sempre empregado no sentido de texto escrito com a utilização do sistema lingüístico da língua inglesa.

[2] A expressão *situation model* é traduzida para o português como *modelo situacional*. Segundo Kintch, modelo situacional é um modelo mental da situação descrita no texto; isto é, uma representação mental do texto, que o leitor constrói, e que requer uma integração entre a informação do texto, o conhecimento prévio do leitor e os seus objetivos de leitura. A explicação da expressão é dada pelo autor nos seguintes termos: "the reader must construct a *situation model* – a mental model of the situation described by the text." (p.73)

[3] "The goal of text comprehension vary widely. The reason for reading a manual might be to learn how to perform an action; one might read a detective story to be entertained for a while; one reads the newspaper to be informed, and so on. But in educational contexts, the goal is often to learn from a text, that is, to construct a situation model that will be remembered and can be used effectively when the information provided by that text is needed in some way at a later time."

estudantes tem conhecimento prévio mínimo sobre ciências."[4]. A dificuldade de entendimento de textos científicos também pode ser devida à falta de domínio do leitor do tipo de registro utilizado neste tipo de texto. Como todo gênero de texto, o científico também tem suas características particulares, que se não fizerem parte dos conhecimentos do leitor podem interferir de forma prejudicial na sua leitura e compreensão.

Observando alguns tipos de avaliação de leitura, como o AVA, o PISA, o SAEB e o NAAL, percebi uma preocupação recorrente com a busca de maior qualidade dos itens que compõem o teste e deste como um todo, bem como com a diminuição das imperfeições no sistema de correção, visando minimizar os efeitos negativos de interferências externas na análise dos resultados obtidos. Nesta busca esses testes lançam mão, entre outras coisas, de modelos de análise estatísticos, e estabelecem uma escala de classificação específica para cada tipo de teste, que parece funcionar como um facilitador da análise e comparação de todo o processo avaliativo, ao longo do tempo.

Porém, apesar de todos esses estudos, de todo o progresso já realizado nesta área e da busca da utilização de formas mais adequadas de medição, por alguns segmentos educacionais, na prática o que se pode perceber é que uma porcentagem muito pequena dos profissionais que estão diariamente elaborando e aplicando avaliações para os mais diferentes propósitos têm o conhecimento teórico necessário para embasar a elaboração e correção de suas avaliações, conferindo-lhes a validade e confiabilidade desejadas, isto é, fazendo com que de fato meçam aquilo que pretendem medir e gerem resultados analisados de forma justa, adequada e confiável dentro dos propósitos a que se destinam.

Durante o desenvolvimento da minha pesquisa "Avaliação de Leitura em Inglês como Língua Estrangeira no Centro de Línguas da UFPR – CELIN", monografia de conclusão do Curso de Bacharelado em Inglês na UFPR (2005), a análise curricular do curso de Letras da UFPR e entrevistas feitas com alunos da universidade, demonstraram que os graduandos de Letras da referida

[4] "The problems are especially important for readers with poor scientific knowledge. In fact, all of the difficulties are exacerbated by the fact that most students have minimal background knowledge about science"

universidade, de maneira geral, não recebem instruções específicas a respeito de elaboração e análise de testes. A entrevista feita então com alguns professores do Centro de Línguas e Interculturalidade – CELIN, que eram alunos de graduação em Letras da UFPR, mostrou que em grande parte os testes eram elaborados e analisados de forma intuitiva, com base em testes pré-existentes e contando com ajuda dos colegas profissionais mais experientes[5].

A utilização de modelos estatísticos de análise de testes ainda é muito reduzida no dia-a-dia acadêmico. De maneira geral, os processos de elaboração e de análise das avaliações como as de suficiência em LE no meio acadêmico assemelham-se ao processo encontrado e relatado na pesquisa realizada no CELIN, anteriormente citada. Isto é, esse processo toma como base muito mais a intuição e a experiência do profissional, que o conhecimento teórico específico sobre o assunto.

Devido a essas considerações acima e à minha experiência como docente de cursos de leitura instrumental da língua inglesa, o foco desta pesquisa são as avaliações de suficiência de leitura em inglês para fins acadêmicos, na tentativa de buscar meios que possibilitem uma elaboração e correção desses testes mais válida e confiável, tanto do ponto de vista da instituição quanto dos candidatos ao ingresso em cursos de mestrado e doutorado.

1.1 OBJETIVOS DE PESQUISA

É evidente a necessidade de utilização de alguma forma de avaliar a capacidade de leitura em língua estrangeira dos alunos dos cursos de pós-graduação, em função da necessidade que a instituição tem de garantir que estes alunos serão capazes de buscar por si mesmos as informações

[5] Neste caso, refiro-me especificamente aos testes de leitura em inglês, foco da minha pesquisa de bacharelado. Porém, os mesmos entrevistados relataram, informalmente, que isso se aplicava às outras habilidades também.

necessárias às suas pesquisas e que serão capazes de utilizá-las da forma adequada, com base em um entendimento adequado dos textos lidos.

Baseada na forma como as avaliações de leitura deveriam ser elaboradas, de acordo com os estudos sobre avaliações em geral, sobre leitura compreensiva e especificamente sobre avaliações de leitura e ainda na necessidade bastante presente de se conferir características consideradas fundamentais aos testes, tais como as chamadas validade e confiabilidade, esta pesquisa tem a intenção de verificar na prática a viabilidade da utilização de métodos teóricos de elaboração e correção de testes e métodos estatísticos de análise de resultados de testes, nas avaliações de suficiência de leitura em inglês como LE para mestrado e doutorado.

1.1.1 Objetivo Geral

Usando como base a teoria sobre elaboração e aplicação de avaliações e o Modelo Rasch de análise estatística de dados, esta pesquisa tem como objetivo geral a elaboração e análise de um teste de suficiência de leitura em inglês para os cursos de mestrado e doutorado, observando a utilidade prática de modelos estatísticos como este e a possibilidade de utilização do teste elaborado para posterior equalização de outros testes de suficiência.

1.1.2 Objetivo Específico

Para que o objetivo geral possa ser alcançado, foram traçados objetivos específicos que viabilizem a realização de tal objetivo. São eles:

a) Realizar um levantamento bibliográfico sobre leitura e avaliação;

b) Realizar um levantamento teórico com vistas a embasar a utilização e interpretação dos dados gerados pelo Modelo Rasch de análise estatística de avaliações.
c) Elaborar e aplicar um teste piloto de suficiência para os cursos de pós-graduação com base no levantamento bibliográfico de avaliações, que possibilite o reconhecimento e correção dos problemas apresentados pelo teste, tanto relacionados à elaboração quanto à correção.
d) Aplicar o teste final, elaborado a partir do piloto, fazer a correção de acordo com o levantamento bibliográfico relativo a correção de testes.
e) Fazer uma análise crítica dos resultados obtidos na aplicação e correção do teste piloto e do teste final, com base no levantamento teórico e com a utilização do aplicativo de análise estatística de dados (Modelo Rasch).

1.2 JUSTIFICATIVA

Partindo da pesquisa e da experiência como docente como mencionado anteriormente percebi que as avaliações em geral são, de fato, muito intuitivas, baseadas no bom senso de cada professor e não raras vezes, na falta dele, infelizmente, graças à qualidade duvidosa de algumas instituições de ensino universitário. Sob o ponto de vista do candidato, não é preciso questionar muito sobre a justiça dessa situação para se obter a resposta, que parece um tanto obvia. Na posição de avaliados, todos se submetem a testes com o intuito de saber ou comprovar a extensão do seu conhecimento em determinada área do conhecimento, em determinada habilidade de linguagem, por exemplo, e sendo assim, gostariam que todo teste refletisse com precisão aquilo que eles sabem ou que são capazes de fazer e que a análise feita pelo avaliador fosse justa e o mais isenta de subjetividade possível.

A responsabilidade do avaliador, neste caso, do profissional responsável pela elaboração, aplicação e correção dos testes, é muito grande, pois o grau de confiança que pode ser depositado no resultado obtido depende da qualidade do seu trabalho. A importância de se dedicar aos estudos sobre testes vem sendo cada vez mais considerada entre os estudiosos, em função da validação do teste. MacNamara (2000, p.48) sustenta que "O propósito da validação no teste de língua é assegurar a defensibilidade e justiça das interpretações baseadas no desempenho do teste." Segundo ele ainda:

> Há muitas razões para se desenvolver uma compreensão crítica dos princípios e práticas de avaliação de língua. [...] Primeiro, testes de língua representam um papel preponderante na vida de muitas pessoas, agindo como passaporte para importantes momentos de transição na escola, no trabalho, e em viagens internacionais. Como testes de língua são mecanismos para o controle institucional dos indivíduos, é evidentemente importante que eles devam ser entendidos, e que passem por um exame minucioso. Em segundo lugar, você pode estar trabalhando com testes de língua em sua vida profissional como um professor ou administrador do teste, preparando alguém para um teste, administrando testes, ou baseando-se em informações dos testes para tomar decisões sobre o nivelamento dos alunos para determinados cursos. Finalmente, se você está conduzindo uma pesquisa sobre o estudo da língua você pode precisar ter medidas da proficiência de língua dos seus sujeitos. [6] (MCNAMARA 2000, p.4-5, tradução nossa)

A argumentação de McNamara demonstra com clareza o poder que os testes dão àqueles que os utilizam. Esse poder acarreta como conseqüências naturais a responsabilidade com a justiça e o comprometimento com a qualidade dos testes, que por sua vez, levam diretamente à necessidade de expandir os conhecimentos sobre o assunto.

---

[6] There are many reasons for developing a critical understanding of the principles and practices of language assessment. […] First, language tests play a powerful role in many people's lives, acting as gateways at important transitional moments in education, in employment, and in moving from one country to another. Since language tests are devices for the institutional control of individuals, it is clearly important that they should be understood, and subjected to scrutiny. Secondly, you may be working with language tests in your professional life as a teacher or administrator, teaching to a test, administering tests, or relying on information from tests to make decisions on the placement of students on particular courses. Finally, if you are conducting research in language study you may need to have measures of the language proficiency of your subjects.

Sob o ponto de vista de um pesquisador, deve-se considerar o fato de que o trabalho realizado com o conhecimento de seres humanos – inteligência, habilidade, capacidade, etc – está focalizado em assuntos que envolvem características necessariamente subjetivas. Não se pode fugir dessa realidade, mas certamente existem meios de se atenuar os efeitos que essa subjetividade exerce sobre as análises e julgamentos que são feitos.

A subjetividade que envolve avaliações psicológicas e de conhecimento como as supracitadas podem levar alguns profissionais da área ao questionamento da necessidade de um trabalho de avaliação de leitura em LE como esse, considerando o fato de que se está analisando algo muito subjetivo e sujeito à influência de vários fatores externos à avaliação em si, e que, por isso, esse estudo talvez não se justifique, devido à impossibilidade de obtenção de análises objetivas e precisas dos dados.

Há que se reconhecer que por mais precisas que sejam quaisquer análises de dados subjetivos, como é o caso da compreensão de leitura, os resultados não serão a representação fiel da realidade, mas uma estimativa do que provavelmente seja o dado real, ao qual não se pode ter acesso direto. No entanto, apesar, e mesmo devido a essa imprecisão na análise dos dados, cabe aos profissionais da área de ciências humanas a realização de mais esforços para que essas análises possam ser consideradas cada vez mais confiáveis e válidas. Isso será possível com a apresentação de resultados que gerem menos desconfiança quanto à imparcialidade e variabilidade do corretor, ou quanto a estarem abrangendo os assuntos que de fato são relevantes ao processo, por exemplo.

Vários profissionais também podem argumentar, não sem razão, que é preciso considerar, também, a existência de diversos fatores que fogem ao controle dos avaliadores e que influenciam nos resultados dos testes, como humor, problemas de saúde ou de stress do candidato no momento da avaliação, entre outros tantos. No entanto, esses fatores não inviabilizam a busca de uma forma de se fazer julgamentos e tomar decisões baseados em resultados de testes, que reflitam com maior precisão os dados analisados e que, além disso, ofereçam ao indivíduo testado resultados baseados em razões mais teóricas e objetivas e menos intuitivas e subjetivas.

Além disso, os profissionais mais críticos podem ainda alegar que avaliações como as supracitadas, que utilizam métodos mais objetivos de análise, com o uso de modelos estatístico, são realizadas com a tomada de uma grande quantidade de dados e, talvez, o relativamente pequeno número de candidatos à pós-graduação não justifique a utilização de um processo de análise mais trabalhoso como este que está sendo proposto.

Embora o número de candidatos seja relativamente pequeno, deve-se ressaltar que, apesar de trabalhosa, a montagem de uma escala de classificação e a utilização de um modelo estatístico na análise, além de facilitar o trabalho a médio e longo prazo, confia às tomadas de decisões maior confiabilidade, de ambas as partes, instituição e candidatos.

Os objetivos de uma pesquisa sobre avaliação de leitura em LE como essa é justamente buscar meios de proporcionar ao teste maior confiabilidade e validade; possibilitar ao profissional acesso a subsídios mais precisos para a elaboração, correção e análise de testes, de forma a facilitar e contribuir com a sua autoconfiança na execução do trabalho e possibilitar que o indivíduo testado identifique uma qualidade de análise de resultados mais objetiva e justa no processo de avaliação como um todo.

Finalmente, vale à pena retornar à pergunta sobre qual a necessidade prática de se testar a compreensão de leitura em LE, em candidatos a cursos de pós-graduação; necessidade essa que justifique tal pesquisa. Dada a natureza e o objetivo desses cursos, parece ser um consenso entre as instituições de ensino superior a necessidade de que o candidato ao curso seja capaz de acessar, organizar, relatar e produzir conhecimentos dentro da área escolhida, que será apresentada na forma de dissertação ou tese. É fato que para tal, a demanda de leitura a ser realizada durante o processo é bastante grande, tanto em língua materna, daqui em diante referida como LM, quanto em LE. Sendo assim, entendo ser fundamental que os candidatos aos cursos de pós-graduação sejam leitores fluentes[7] em uma segunda língua ou, no caso

[7] Neste caso, o conceito de *leitor fluente* depende diretamente dos padrões estabelecidos por cada instituição, para a capacidade de compreensão de leitura considerada suficiente para os propósitos em questão.

do doutorado, em uma terceira, de acordo com as exigências das próprias instituições.

1.3 ORGANIZAÇÃO

Esta dissertação está organizada em seis capítulos, com a abordagem dos seguintes assuntos:

O capítulo I inclui a introdução, os objetivos desta pesquisa, a sua justificativa e a forma como a dissertação está organizada.

O capítulo II contém o levantamento bibliográfico sobre leitura. Deste levantamento, constam conceitos sobre a leitura, o leitor e o texto, considerados relevantes para a pesquisa.

O capítulo III traz o levantamento bibliográfico sobre avaliações, no que se refere aos conceitos gerais sobre avaliação, ao processo avaliativo, à correção de testes, à Teoria Clássica de Testes e também sobre a análise estatística de dados, baseada na Teoria de Resposta ao Item (TRI), na qual é embasado o modelo Rasch, utilizado na análise de dados.

No capítulo IV estão detalhadas: a metodologia empregada na pesquisa, o desenvolvimento dos testes aplicados, a correção dos testes e a análise dos resultados, incluindo alguns conceitos não abordados anteriormente, mas importantes para a compreensão do processo de elaboração e correção dos testes e da análise dos resultados, com base no levantamento bibliográfico e no Modelo Rasch.

No capítulo V são apresentadas as análises dos resultados obtidos, a partir dos gráficos e dados gerados pelo modelo.

Ao final são apresentas a conclusão da pesquisa, as limitações encontradas no estudo e sugestões para futuras pesquisas.

2 REVISÃO BIBLIOGRÁFICA DE LEITURA E AVALIAÇÃO

O objetivo traçado para esta pesquisa envolve o desenvolvimento e análise de testes de leitura em inglês. Para que este intento fosse alcançado, foi necessário, primeiramente, que se fizesse uma reflexão sobre conceitos do que é leitura, do que é compreensão de leitura, a fim de que aqueles conceitos que subjazem a elaboração dos testes pudessem ser esclarecidos. Sendo assim, neste segundo capítulo é apresentada uma revisão bibliográfica sobre leitura, com os conceitos que esta autora considera fundamentais para o embasamento teórico da pesquisa realizada.

Em função da qualidade do ensino em geral, no Brasil, o que se pode observar nas escolas, quer de primeiro e de segundo graus, quer em universidades, são alunos com grandes dificuldades em compreender aquilo que lêem e, conseqüentemente, de incorporar os novos conhecimentos encontrados, pelo fato de não conseguirem fazer a relação adequada entre o conhecimento prévio e o novo, ou de não conseguirem apreender os julgamentos de valor expostos no texto, por exemplo. Isso se deve, em parte, à concepção errônea de que a leitura é uma das habilidades mais fáceis de se aprender e utilizar, talvez pelo fato de ser, em grande medida, uma habilidade receptiva. No entanto, esta habilidade envolve elementos bastante complexos.

Jouve inicia o primeiro capítulo de seu livro declarando a complexidade e a existência de várias facetas na leitura. Afirma ele que

> A leitura é antes de mais nada um ato concreto, observável, que recorre a faculdades definidas do ser humano. Com efeito, nenhuma leitura é possível sem um funcionamento do aparelho visual e de diferente funções do cérebro. Ler é, anteriormente a qualquer análise do conteúdo, uma operação de percepção, de identificação e de memorização dos signos. [...] Assim, considerada nos seu aspecto físico, a leitura apresenta-se, pois como uma atividade de antecipação, de estruturação e de interpretação. (2002, p.17).

Mas, a leitura é muito mais que um processo físico, ela é um processo mental, que envolve a relações de vários elementos. Jouve (2002) afirma que a leitura é, além de um processo físico, um processo cognitivo, em que, após

identificar e decifrar os signos, o leitor procura entender o que eles significam, e que a leitura é um processo:

a) *afetivo*, na medida em que as informações contidas no texto influem de alguma forma nas emoções do leitor;
b) *argumentativo*, uma vez que todo texto é sempre passível de algum tipo de análise e;
c) *simbólico*, porque o sentido que se pode construir a partir do texto depende do contexto sociocultural e lingüístico e do propósito de leitura de cada leitor.

Todas essas idéias constituem um panorama geral do que seja a leitura, que serve de ponto de partida para o entendimento do que vamos denominar, neste momento, *compreensão de leitura*, mas que poderá também ser tratada, indistintamente ao longo deste trabalho, como *compreensão de texto, compreensão* ou simplesmente *leitura*.

Para que a importância de se definir a compreensão de leitura seja observada, é preciso lembrar para que ela serve, onde ela é usada, porque é necessário compreender um texto que se lê de forma adequada, e que forma adequada é essa. A esse respeito Snow lembra que

> contribuintes e empregadores concebem a compreensão de leitura como uma das habilidades que estudantes com ensino médio completo deveriam ter adquirido durante seus anos na escola. O corpo docente das universidades considera que níveis altos de compreensão de leitura são pré-requisito para o sucesso do aluno."[8] (2002, p.10, tradução nossa)

Pode-se perceber por meio do exemplo acima, que podem existir diversas concepções do que seja compreensão de leitura, dependendo do contexto em que ela está inserida. É isso que será apresentado a partir de agora.

---

[8] "Taxpayers and employers think of reading comprehension as one of the capabilities that high school graduates should have acquired during their years in school. University faculty view high levels of reading comprehension as a prerequisite to a student's success."

2.1 O QUE É COMPREENSÃO DE LEITURA

Uma forma inicial de perceber a complexidade e a abrangência que envolve o conceito de compreensão de leitura é observar um exemplo simples do que pode ser considerada uma forma de compreensão de leitura. A decodificação de uma placa de trânsito em um cruzamento, onde está escrita apenas a palavra PARE é também uma compreensão de texto. Neste caso o leitor precisa identificar elementos tais como o código lingüístico utilizado e o contexto em que ele está sendo utilizado, isto é, o código de trânsito, para compreender que o que está implícito neste texto é a indicação de que o motorista deve parar o veículo que está conduzindo naquele lugar indicado e verificar a possibilidade de cruzar a via a sua frente. Essa leitura e compreensão devem ser, e são feitas por todo motorista em frações de segundos. Mas os tipos de leitura são muito variados e de acordo com eles existem muitas formas de expressar o conceito do que seja a compreensão de leitura.

Duke (2004, p.93) afirma que o que comumente denomina-se *compreensão* envolve processos que variam de acordo com o texto, o tópico e o propósito da leitura. Neste trabalho serão adotados alguns daqueles conceitos de leitura que defendem o ponto de vista escolhido para esta pesquisa. O sentido de compreensão de leitura que motivou o desenvolvimento deste trabalho de avaliação de leitura em inglês se reflete no conceito adotado por vários autores, alguns dos quais serão apresentados a seguir.

De acordo com Foucambert

> Ler significa ser questionado pelo mundo e por si mesmo, significa que certas respostas podem ser encontradas na escrita, significa poder ter acesso a essa escrita, significa construir uma resposta que integra parte das novas informações ao que já se é. (1994, p.5).

Cohen (1994, p.212) por sua vez, afirma que “Ler requer que o leitor supra o significado ativamente com regularidade.” (tradução nossa). E, para Carrell,

> não apenas o leitor é um participante ativo no processo de leitura, fazendo predições e processando informação, mas tudo na experiência ou formação de conhecimento prévio do leitor representa um papel potencial no processo. (1987, p. 24, tradução nossa).

Pode-se entender que, sob o ponto de vista desses autores, a leitura é um processo ativo e não passivo; ela é mais que atribuição de significado, é construção de sentido. Aparentemente, isso faz com que o leitor passe a ocupar o papel principal no processo de leitura, uma vez que, sem ele não há construção de sentidos possível, e conseqüentemente, razão para o próprio texto, visto que é a partir do leitor que um texto tem sua existência confirmada. No entanto, é preciso considerar que, a partir do momento em que o texto é escrito, ele existe, mesmo antes que um leitor o leia, e continuará a existir, mesmo que esse leitor nunca tome conhecimento de sua existência. A partir dos pontos de vista supracitados, compreende-se que, na medida em que todo texto pressupõe certos conhecimentos prévios, necessários na construção de novos sentidos, toda leitura que se faz de um texto é, de certa forma, singular, pois cada leitor traz em si uma gama de conhecimentos única, diferenciada ao longo de sua história de vida, não só enquanto leitor, mas enquanto indivíduo. Entretanto, se cada leitura fosse considerada única, ela permitiria interpretações que poderiam extrapolar o que o texto propõe e o que ele permite lingüisticamente. Além disso, como conseqüência dessa visão de leituras únicas, a compreensão de leitura não poderia ser avaliada, uma vez que, em sua singularidade, ela não poderia ser contestada. Pode-se perceber, então, que o leitor vai suprir os significados, participar ativamente no processo e construir novos significados dentro de certos limites.

Na prática, é como se o leitor fizesse um “diálogo” entre o que ele já sabe e o que está lendo, criticando, analisando, aceitando, recusando as informações e idéias do texto, por exemplo. Dessa forma, ele reorganiza o seu próprio conhecimento e constrói novas idéias e conceitos, ampliando seus conhecimentos. Nesse processo, quanto maior o conhecimento prévio do leitor sobre o assunto, mais fácil se torna a compreensão de um texto. Da mesma forma, quanto mais complexa é a abordagem de um assunto em um texto, mais conhecimento prévio se faz necessário, para uma adequada compreensão do mesmo. Entretanto, a utilização do conhecimento pressuposto pelo texto e as

construções de sentido que o leitor pode fazer a partir do texto têm uma limitação imposta pelo código lingüístico utilizado, pelos elementos lingüísticos que compõem o texto e pelo contexto em que o texto é lido, por exemplo. Caso essas limitações não se impusessem, obviamente os textos se tornariam inviáveis, com o passar do tempo, dado o volume de informações, sempre crescente, que precisariam conter e a variação de sentidos apreendidos, que por sua vez, impossibilitaria que se chegasse a um sentido geral qualquer que pudesse ser partilhado pela maioria dos leitores, uma vez que cada leitor poderia inferir qualquer coisa que lhe aprouvesse e que, em sua opinião fosse o melhor sentido.

Corroborando essas afirmativas, Kintsch e Kintsch (2004, p.71) afirmam que a compreensão de leitura não é um processo unitário, mas um processo que envolve vários componentes, que integram o que está sendo lido com o conhecimento prévio e a experiência de quem está lendo, sendo que essa integração está sujeita aos limites impostos pelo contexto que envolve e determina o processo da leitura. Koch e Elias apontam que na perspectiva interacional, ou dialógica, da língua, em que existe a interação entre autor – texto – leitor,

> o **sentido de um texto é construído na interação texto-sujeitos** e não algo que preexista a essa interação. **A leitura** é, pois, uma atividade interativa altamente complexa de produção de sentidos, que **se realiza** evidentemente **com base nos elementos lingüísticos presentes na superfície textual e na sua forma de organização**, mas requer a mobilização de um vasto conjunto de saberes no interior do evento comunicativo. (2006, p. 11, grifo nosso).

Apontando a influencia dos elementos lingüísticos do texto ainda, segundo Snow, o Reading Study Group (RAND) define o termo

> compreensão de leitura como o processo de extrair e construir significados simultaneamente através da interação e envolvimento com a linguagem escrita. Ela consiste de três elementos: o leitor, o texto, e a atividade ou propósito para a leitura.[9] (2002, p.xiii, tradução e grifo nosso).

[9] *"reading compreension* as the process of simultaneously extracting and constructing meaning through interaction and involvement with written language. It consists of three elements: the reader, the text, and the activity or purpose for reading."

De acordo com esta autora, estes três elementos, ou dimensões, definem “um fenômeno que ocorre dentro de um contexto sócio-cultural maior que determina e é determinado pelo leitor e que interage com cada um dos três elementos.”[10] A autora representa essa relação através da figura abaixo.

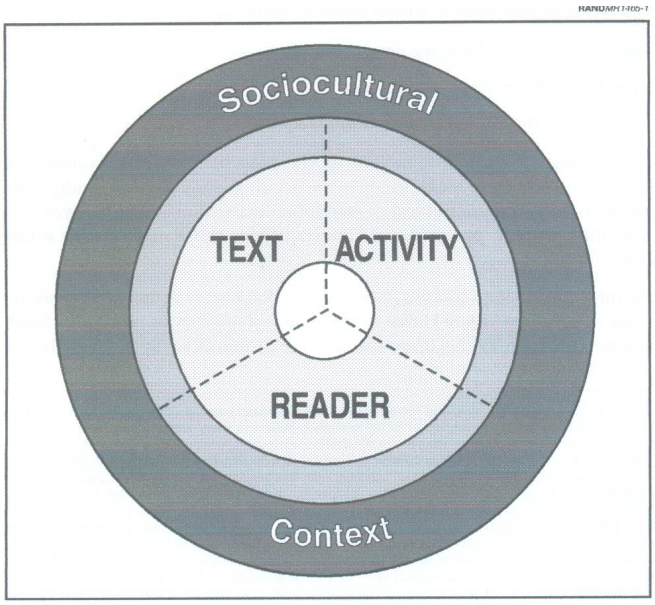


Figure 2.1—A Heuristic for Thinking About Reading Comprehension

FIGURA 1 – A HEURISTIC FOR THINKING ABOUT READING COMPREHENSION.
FONTE: *READING FOR UNDERSTANDING* (2002)

É na interação dos três elementos acima que se dá o processamento textual, ou seja, o processamento das informações relevantes dentro daquele contexto sociocultural determinado. Cabe ao leitor determinar como se dará a interação entre os três elementos dentro desse contexto previamente estabelecido, a fim de que ele consiga atingir os objetivos propostos. Ou seja, o leitor deverá decidir que recursos ele utilizará para compreender o texto, ou as partes das quais se valerá para a realização da tarefa proposta. Em outras palavras, se ele vai fazer uma leitura rápida do texto para saber qual é a idéia geral apresentada; se ele vai ler detalhadamente para apreender o máximo de informações possível ou se ele vai apenas procurar uma informação bem específica dentro do texto. E ainda, se ele precisa saber o significado exato de

[10] “a phenomenon that occurs within a larger *sociocultural context* that shapes and is shaped by the reader and that interacts with each of the three elements.”

alguma palavra ou sentença, e neste caso, decidir se vai usar o contexto lingüístico da palavra ou sentença para apreender o significado, ou vai fazer uso de um dicionário, por exemplo.

Diferentes autores enfatizam o uso de diferentes processos de leitura como relevantes durante o processamento textual. Kato afirma que estudiosos das áreas de ciência da cognição e inteligência artificial

> referem-se a dois tipos básicos de processamento de informação: o que chamam de *top-down* (literalmente = descendentes) e o que chamam de *bottom-up* (literalmente = ascendente).O processamento descendente (*top-down*) é uma abordagem não-linear, que faz uso intensivo e dedutivo de informações não-visuais e cuja direção é da macro para a microestrutura e da função para a forma.O processamento ascendente (*bottom-up*) faz uso linear e indutivo das informações visuais, lingüísticas, e sua abordagem é composicional, isto é, constrói o significado através da análise e síntese do significado das partes. (2007, p. 50).

A respeito desses dois tipos de processamento, Nuttall explica que atualmente entende-se que

> na prática um leitor muda continuamente de um foco para outro, ora adotando uma abordagem *top-down* para predizer um provável significado, depois mudando para a abordagem *bottom-up* para verificar se aquilo é realmente o que o escritor está dizendo.[11] (2000, p. 17, tradução nossa)

O uso complementar dessas abordagens é chamado de *leitura interativa* (literalmente: *interactive reading*), e, segundo ela, as duas abordagens "podem ser ativadas por escolha consciente, e ambas são importantes estratégias para os leitores".[12] (tradução nossa)

Koch e Elias (2006, p.12-13) adotam a idéia de que a leitura é uma atividade de interação entre o autor, o leitor e o texto, em que o leitor é um construtor de sentido. E nesta atividade de produção de sentidos o leitor utiliza certas estratégias, como:

---

[11] "in practice a reader continually shifts from one focus to another, now adopting a top-down approach to predict the probable meaning, then moving to the bottom-up approach to check whether that is really what the writer says."

[12] "Both approaches "can be mobilized by conscious choice, and both are important strategies for readers."

a) antecipações;
b) hipóteses;
c) inferências;
d) verificações.

Nuttall (2000) relaciona entre diversos processos utilizados na leitura, dois que, segundo ela, permitem que o leitor selecione as partes do texto que são mais relevantes e que precisam de uma leitura mais detalhada, conforme os propósitos de leitura estabelecidos. São elas:

a) ler por alto (denominado *skimming*), que o leitor usa para determinar a idéia geral do texto, ou de um parágrafo, por exemplo.
b) passar os olhos em um texto (denominado *scanning*), que o leitor pode utilizar para encontrar informações específicas, ou para ter uma primeira impressão do texto a ser lido.

Além desses processos de leitura, existem outros relacionados ao conhecimento de mundo de que o leitor se utiliza na compreensão de textos. Em um dos estudos apoiado pelo Departamento de Educação dos Estados Unidos, realizado pelo programa "Regional Educational Laboratory" (REL), no Pacific Resources for Education and Learning (PREL), (2005, p.9), o foco é a compreensão de leitura. Segundo este estudo, um bom leitor, quando está lendo, ativa determinadas redes de conhecimentos, ou esquemas (*schema),* os quais estão relacionadas ao tópico ou às palavras que encontra no texto e utiliza processos cognitivos e metacognitivos durante a leitura. Os termos cognição e metacognição são definidos neste estudo da seguinte forma: "Cognição refere-se a funções mentais tais como lembrar, focalizar a atenção, e processar informação. *Metacognição* refere-se à consciência do leitor da sua cognição; isto é, sua reflexão sobre sua reflexão."[13] (tradução nossa)

[13] "*Cognition* refers to mental functions such as remembering, focusing attention, and processing information. *Metacognition* refers to readers' awareness of their cognition; that is, their thinking about their thinking."

Em Kato (2007, p. 124), lê-se que uma das categorias de conhecimento metacognitivo propostas por Flavell (1978)[14], e com a qual ela trabalha em seu livro, considera que este conhecimento "controla e seleciona aspectos da atividade cognitiva". Kato propõe que as estratégias *cognitivas* de leitura designam "os princípios que regem o comportamento automático e inconsciente do leitor" e as estratégias *metacognitivas* de leitura designam "os princípios que regulam a desautomatização consciente das estratégias cognitivas."[15]

Bloom (1956), citado por Graesser, León e Otero (2002, p. 09-10) propõe uma tabela dos principais tipos de processos cognitivos. Essa tabela, no entanto, de acordo com os autores acima, não abrange todas as distinções teóricas feitas pelos pesquisadores atualmente sobre esse assunto. Mesmo assim, esta tabela, traduzida abaixo, apresenta uma tipologia, que para efeitos desta pesquisa, é suficiente, por ora, para a compreensão geral sobre processos cognitivos e metacognitivos.

Tipos de Processos Cognitivos	
Reconhecimento	O processo de identificação de verbete de conteúdo específico (e.g., termos, fatos, regras, métodos, princípios, procedimentos, objetos) que foram mencionados explicitamente no texto.
Recordação.	O processo de resgatar ativamente da memória e produzir de conteúdo que foi mencionado explicitamente no texto.
Compreensão.	Demonstração de entendimento do texto em nível de modelo mental[16] através da geração de inferências, e interpretação, paráfrase, tradução, explicação, ou resumo de informação.

[14] De acordo com Kato(2007): "J Flavell (1978) "Cognitive Monitoring". Draft prepared for the conference on Children's Oral Communication Skills, U. of Wisconsin. *Apud* Moore (1983)."

[15] Para maiores esclarecimentos sobre os conceitos de cognição e metacognição propostos por Kato ver capítulo 8 da obra constante na indicação bibliográfica.

[16] Ver explicação do termo na nota de rodapé número 2 ou no glossário.

Aplicação.	O processo de aplicar o conhecimento extraído do texto a um problema, situação ou caso (ficcional ou real) que não foi explicitamente mencionado no texto.
Análise.	O processo de decompor os elementos e fazer ligações entre eles.
Síntese.	O processo de reunir novos padrões e estruturas, tais como a construção de uma nova solução para um problema ou a criação de uma nova mensagem para um público
Avaliação.	O processo de julgar o valor ou eficácia de um processo, procedimento, ou entidade, de acordo com alguns critérios e padrões.

TABELA 1. - TIPOS DE PROCESSOS COGNITIVOS[17].
FONTE: BLOOM APUD GRAESSER, LEÓN E OTERO (2002)

É importante lembrar que esses processos cognitivos se realizam através de funções mentais automáticas, utilizadas de forma inconsciente por parte do leitor. A partir destes processos podem-se deduzir, então, os processos metacognitivos, que se realizam por meio dessas mesmas funções mentais, que são selecionadas e utilizadas pelo leitor de forma consciente na leitura na medida em que cada uma delas se faz necessária.

A utilização dessas funções mentais e processos cognitivos no processo de construção de sentido é rápida, flexível e simultânea, o que significa que, consciente ou inconscientemente, o leitor pode usar uma ou mais funções ou processos, conforme a necessidade em cada momento da leitura. Isso acontece, é claro, desde que o leitor tenha pleno domínio sobre os elementos lingüísticos presentes no texto e sobre o assunto abordado, para que não tenha que processar esses recursos de forma controlada.

Koch e Elias (2006, p. 30-40) abordam a questão de que, durante o processamento textual, para que haja construção de sentido a partir da

[17] Ver reprodução da tabela original no anexo 1.

interação autor-texto-leitor, entram em jogo, ainda, os conhecimentos do leitor, que recorre a três grandes sistemas de conhecimento:

a) conhecimento lingüístico - abrange o conhecimento lexical e gramatical;
b) conhecimento enciclopédico – conhecimentos gerais ou conhecimento de mundo;
c) conhecimento interacional – formas de interação por meio da linguagem, como por exemplo os objetivos ou propósitos do autor; quantidade de informação necessária e seleção de variante lingüística; sinais de articulação do texto e elementos de apoio textual; identificação do gênero textual e da estrutura do texto.

São estes conhecimentos que permitem que o leitor faça o processamento do texto, que se realiza através da relação entre os elementos presentes no texto; entre as partes de um mesmo texto; entre os elementos ou partes de um texto e o texto na sua integralidade; entre esse e outros textos diferentes e entre o conhecimento que o leitor traz para a leitura e as informações novas que ali ele encontra.

No que diz respeito a esta pesquisa um dos conhecimentos acima é especialmente relevante, por interferir de maneira mais evidente no desempenho do leitor: o conhecimento lingüístico. (ALDERSON, 2000, p. 34) chama a atenção para o fato de que parece óbvio que “se leitores não conhecem a linguagem do texto, então eles vão ter grande dificuldade no processamento do texto.” (tradução nossa) De acordo com este autor, apenas recentemente os estudos sobre o conhecimento retórico e metalingüístico vem sendo aprofundado em LE. Até então, a ênfase dos estudos recaía na sintaxe e no léxico. “Em leitura em segunda língua e língua estrangeira, foi sempre assumido que aprendizes deveriam primeiro adquirir conhecimento de linguagem antes que eles pudessem ler.” (ALDERSON, 2000, p. 36, tradução nossa) Estudos mais recentes demonstram que, de fato, o conhecimento de certas estruturas sintáticas, ou da habilidade de processá-las é importante para a compreensão de leitura, no entanto, de maneira geral, “um conhecimento do léxico do texto, assim como conhecimento do conteúdo específico e geral, podem muito bem compensar a falta de conhecimento lingüístico.” (ibidem, p.

37, tradução nossa) Além disso, havia também a crença de que se o leitor não fosse um bom leitor em LM ele não seria também um bom leitor em LE e vice-versa. Alderson conclui que os estudos realizados a esse respeito demonstram que o conhecimento da LE é mais importante que a habilidade de leitura em LM. Ele afirma que:

> Pesquisas para investigar ou resolver a questão de se a leitura em segunda língua é um problema de *linguagem* ou um problema de *leitura* sugeriram a idéia de um limiar de conhecimento lingüístico, sem o qual leitores não podem esperar transferir nenhuma habilidade de leitura da primeira língua para a segunda língua. [...] A conclusão clara de tais estudos é que conhecimento da segunda língua é mais importante que habilidades de leitura em primeira língua, e que um limiar lingüístico que existe deve ser transposto antes que a habilidade de leitura na primeira língua possa ser transferida para o contexto de leitura na segunda língua. [...] quanto mais exigente a tarefa, mais alto é o limiar lingüístico. O que torna uma tarefa exigente vai estar relacionado a assuntos como tópico do texto, linguagem do texto, conhecimento prévio e tipo de tarefa. (2000, p. 38-9, tradução nossa).[18]

O que se espera do leitor alvo desta pesquisa é que ele tenha conhecimento da língua inglesa o suficiente para ser capaz de transpor o limiar lingüístico mínimo necessário a um entendimento satisfatório do texto.[19] Porém, o conhecimento lingüístico apenas não basta. Como mencionado anteriormente, para que a compreensão de um texto seja possível, é preciso haver a interação de três elementos: o leitor, com todos seus conhecimentos; a atividade, isto é, o propósito de leitura que motiva a seleção e uso de determinados processos em detrimento de outros e o próprio texto, em que são

---

[18] "Research to investigate or resolve the question whether second-language reading is a *language* problem or a *reading* problem has suggested the notion of a threshold of linguistic knowledge, without which readers cannot expect any first-language reading ability to transfer to the second language. […] The clear conclusion of such studies is that second-language knowledge is more important than first-language reading abilities, and that a linguistic threshold exists which must be crossed before first-language reading ability can transfer to the second-language reading context. […] the more demanding the task, the higher the linguistic threshold. What makes a task demanding will relate to issues like text topic, text language, background knowledge and task type."

[19] As questões do que deve ser entendido por "entendimento satisfatório" e qual é o mínimo necessário estão definidas no item que trata da definição do construto do teste.

consideradas características tais como gênero, código lingüístico, entre outros elementos. E entre os três elementos citados, é preciso estabelecer, neste momento, quem é o leitor visado nesta pesquisa, para que se possa determinar efetivamente o que se espera dele em termos de desempenho de leitura.[20]

2.2 LEITOR IDEAL / LEITOR REAL

Considerando que o leitor é uma das partes integrantes do processo de compreensão de texto, conforme exposto no item acima, é preciso conhecê-lo melhor, para saber o que se pode esperar dele no contexto em que esta pesquisa está inserida, afinal, na atividade de compreensão de leitura a habilidade de leitura varia conforme o leitor e depende das competências, experiências e conhecimentos de cada um.

As definições do que se considera um “bom leitor”, “leitor competente”, “leitor proficiente” pode variar de acordo com a visão que se tem de leitura e com os propósitos estabelecidos para essa leitura. Um bom leitor em LM não será, necessariamente, um bom leitor em LE, se ele não tiver domínio suficiente do código lingüístico utilizado no texto. Da mesma forma, um leitor considerado competente, de maneira geral, pode não ser tão competente assim, em se tratando de um assunto que lhe é desconhecido. E um leitor proficiente de textos literários pode apresentar inúmeras dificuldades em compreender textos técnicos, por exemplo. Existem algumas nuances que precisam ser consideradas na definição do que seja um leitor competente, ou um bom leitor, para efeitos deste trabalho.

A respeito das competências de um leitor, os Parâmetros Curriculares Nacionais afirmam que “um leitor competente sabe selecionar, dentre os textos que circulam socialmente, aqueles que podem atender a suas necessidades, conseguindo estabelecer as estratégias adequadas para abordar tais textos. O

[20] A discussão sobre tipos de texto é apresentada no próximo item deste capítulo e a discussão sobre as atividades (ou seja, os propósitos de leitura) será apresentada nos itens pertinentes à elaboração do teste e das questões.

leitor competente é capaz de ler as entrelinhas, identificando, a partir do que está escrito, elementos implícitos, estabelecendo relações entre o texto e seus conhecimentos prévios ou entre o texto e outros textos já lidos." (PCN, 1998, p. 70) Embora esses parâmetros visem questões de ensino e aprendizagem em LM para crianças e adolescentes, eles podem ser aplicados ao contexto da pesquisa em LE também, abstraindo-se, naturalmente, o fato de que o leitor em LE tem ainda que adicionar a essas habilidades o conhecimento da língua alvo. O que está subentendido na adoção desse conceito para ser utilizado em relação à LE é que as habilidades que determinam o que seja um leitor competente, ou proficiente, vão além do conhecimento da língua, pois são habilidades necessárias a leitores em geral, qualquer que seja a língua usada no texto escrito. Isso significa, em outras palavras, que na compreensão de um texto, além de envolver o que diz respeito ao conhecimento da língua, as competências consideradas nesta pesquisa são requisitos comuns para a compreensão de quaisquer textos, quer em LM, quer em LE.

De acordo com Snow

> Para compreender, um leitor deve ter uma grande variedade de capacidades e habilidades. Isso inclui capacidades cognitivas (por exemplo: atenção, memória, habilidade crítica-analítica, inferência, habilidade de visualização), motivação (um propósito para ler, um interesse no conteúdo sendo lido, capacidade de realização como leitor) e vários tipos de conhecimento (vocabulário, conhecimento do assunto e do tópico, conhecimento lingüístico e de discurso, conhecimento de estratégias específicas de compreensão). E é claro, as capacidades cognitivas, motivacionais e lingüísticas específicas e o conhecimento básico requerido em qualquer ato de compreensão de leitura depende dos textos em uso e da atividade específica na qual alguém está envolvido. (2002, p. 13, tradução nossa).[21]

[21] "To comprehend, a reader must have a wide range of capacities and abilities. These include cognitive capacities (e.g., attention, memory, critical analytic ability, inferencing, visualization ability), motivation (a purpose for reading, an interest in the content being read, self-efficacy as a reader) and various types of knowledge (vocabulary, domain and topic knowledge, linguistic and discourse knowledge, knowledge of specific comprehension strategies). Of course, the specific cognitive, motivational, and linguistic capacities and the knowledge base called on in any act of reading comprehension depend on the texts in use and the specific activity in which one is engaged."

No que concerne ao papel das competências do leitor na compreensão de texto o PREL (2005, p. 7) propõe que se trabalhe com um processo de dois níveis para a leitura: o que eles chamam de *habilidades básicas* (foundational skills) e *processos de leitura de ordem superior* (higher order reading processes). No primeiro nível são aplicadas habilidades "tais como reconhecimento e decodificação de palavras, fluência, e conhecimento de vocabulário" e no segundo, estão envolvidos "os procedimentos usados pelos leitores para fazer conexões entre as palavras [...] e para relacionar o conhecimento existente com a informação do texto de maneira a analisar, avaliar, e pensar sobre o significado das frases, parágrafos, e textos inteiros." (tradução nossa) Os dois níveis são fundamentais para uma compreensão eficaz do texto lido, pois um leitor que possua apenas as competências do primeiro nível não será capaz de construir sentidos possíveis, previsíveis e desejáveis em vários tipos de texto que ele poderá encontrar.

Ainda segundo o PREL, bons leitores antes de ler, por exemplo,

> usam seu conhecimento do assunto do texto para pensar sobre e estabelecer propósitos e expectativas para sua leitura. Enquanto lêem, eles pensam se estão entendendo o texto e, se não, o que podem fazer para melhorar sua compreensão. Depois de ler, eles podem pensar sobre o que leram, se gostaram ou aprenderam alguma coisa com isso, e se a sua leitura deu a eles informações que eles poderão usar no futuro.[22] (2005, p.9. tradução nossa).

Embora o foco desta pesquisa não seja o processo de aprendizado da leitura em LE, entender onde o leitor ideal deve estar posicionado ao longo deste processo de aprendizado ajuda a determinar as características que se espera encontrar nos leitores-alvo e a definir o que se pode e deve esperar do seu desempenho em avaliações de leitura em LE.

---

[22] "use their knowledge of the text subject to think about and set purposes and expectations for their reading. As they read, they think about whether they are understanding the text and, if not, what they can do to improve understanding. After reading, they may think about what they read, whether they enjoyed or learned something from it, and whether their reading gave them ideas of information they might use in the future."

Com relação à situação de aprendizado de leitura Snow afirma que o conceito do que seja ler bem varia de acordo com o estágio de desenvolvimento em que o leitor se encontra. Segundo ela

> aprender a ler bem é um processo de desenvolvimento de longo prazo. No ponto final, o leitor adulto proficiente é capaz de ler uma variedade de materiais com facilidade e interesse, é capaz ler por objetivos variados, é capaz de compreender até mesmo quando o material não é nem fácil de entender nem intrinsecamente interessante.[23] (2002, p. xiii, tradução nossa).

Como visto no item 2.1 acima, contextos socioculturais de leitura pré-determinados, impõem certos limites ao tipo de leitura que deverá ser feita. Neste momento, é preciso fazer um parêntese para chamar a atenção para o fato de que se deve ter em mente que, visto ser a abordagem desta pesquisa o ingresso em cursos de mestrado e doutorado, o leitor-alvo, neste caso, deve preencher minimamente os requisitos de: ser graduado em um curso de nível superior e, conseqüentemente, um leitor adulto; ser proficiente em LM e possuir certo grau de proficiência em Língua Inglesa (grau este que será definido oportunamente). A percepção do contexto acima permite que se coloque este leitor no ponto final do seu processo de aprendizado de leitura em inglês, ou próximo a ele, imaginando que fosse possível estabelecer, de fato, um ponto final para o aprendizado de uma habilidade de linguagem, ou qualquer outro aprendizado.

Kato trata a questão da proficiência do leitor relacionada aos processos de leitura que ele usa. Como mencionado anteriormente, ela aponta o fato de que

> na área de compreensão e leitura, onde temos processos inacessíveis à observação direta, tivemos também, até recentemente, duas concepções radicalmente opostas, oposição essa que se manifesta na denominação com que elas são conhecidas hoje: a hipótese ascendente (bottom-up), ou de dependente do texto, e *a*

[23] "learning to read well is a long-term developmental process. At the end point, the proficient adult reader can read a variety of materials with ease and interest, can read for varying purposes, and can read with comprehension even when the material is neither easy nor intrinsically interesting."

> *hipótese descendente (top-down),* ou dependente do leitor[24].(2007, p. 66)

E propõe

> que o leitor proficiente é aquele que faz uso apropriado desses processos, o que o torna um leitor ao mesmo tempo fluente e preciso. As estratégias são determinadas por vários fatores: o grau de novidade do texto, o local do texto, o objetivo da leitura, a motivação para a leitura, etc" (KATO, 2007, p. 68),

ou seja, o leitor proficiente é aquele capaz de utilizar o processo ascendente e descendente citados acima de forma complementar, de acordo com as necessidades identificadas por ele em cada situação. Isso significa que um leitor ideal deveria ser capaz de ler compreensivamente qualquer gênero de texto e utilizar as estratégias de leitura necessárias para realizar seguintes tarefas: reconstituição da informação, em que o leitor deve identificar e extrair as informações solicitadas, da forma como elas aparecem no texto; ordenação e relevância, em que o leitor tem que reconstituir a ordenação das informações apresentadas, considerando o grau de relevância das mesmas; estabelecimento de relações, em que o leitor deve ser capaz de estabelecer as relações entre elementos textuais, partes do texto, o texto todo e outro texto, em todas as combinações possíveis; reconhecimento do quadro enunciativo, que exige compreensão de estratégias discursivas, como reconhecimento do sujeito do discurso, por exemplo; apreensão e julgamento de valor, em que o leitor tem que ser capaz de identificar e entender os julgamentos de valor sobre as informações, que estão presentes no texto e reconstrução da argumentação, que exige a capacidade de reconstrução da linha argumentativa utilizada na apresentação da informação. Esses tipos de tarefas são os seis tipos básicos de questão propostos por Nery (2003), que ainda adiciona a eles outros três pares de categorias, que são: questões pontuais ou globais, lineares ou não lineares, orientadas ou não orientadas.[25]

---

[24] Para maiores informações sobre esses processos consultar, entre outros autores, Nuttall (2000) e Brown (2001) – ver referências bibliográficas.

[25] Ver reprodução do esquema geral da matriz de questões no anexo 2 e classificação de questões no item 4.2.5, que trata especificamente de questões.

Todas as descrições acima dizem respeito a um leitor ideal. Entretanto, é preciso considerar que os leitores ideais são minoria, no contexto sócio-educacional brasileiro atual. Os leitores reais, geralmente não possuem todas as competências citadas e, além disso, as suas habilidades cognitivas e metacognitivas de leitura, tanto em LM quanto em LE, variam consideravelmente, conforme se pode deduzir pela observação dos resultados do PISA (Programme for International Student Assessment), realizado em 2000, e cuja ênfase foi a leitura. De acordo com esta avaliação dos cinco níveis[26] de letramento estabelecidos, o Brasil atingiu, na média geral, o nível 2. Segundo Claudio de Moura e Castro, no relatório do PISA 2000,

> os resultados são decepcionantes. Cinco por cento dos nossos alunos sem atraso conseguem chegar ao nível 4 de compreensão dos textos e somente 1% chega ao nível 5. Compare-se com 31% e 6% para a Coréia, para os mesmos níveis, 22% e 13% para os Estados Unidos (sem atraso) e 21% e 4% para a Espanha (sem atraso). (p. 88).

A razão da decepção pode-se ver na tabela abaixo, em que se observa que apenas 1% dos alunos brasileiros com nove anos ou mais de educação formal conseguiu atingir o nível 5, que é o mais alto. Entre os alunos com oito anos de escolarização nenhum atingiu o nível 5 e apenas 1% atingiu o nível 4. E entre os alunos com sete anos de escolarização nenhum atingiu os dois níveis mais altos e somente 1% conseguiu atingir o nível 3.[27]

Proporção de alunos de diversas séries nos diferentes níveis de proficiência												
Países	Abaixo de 1		Nível 1		Nível 2		Nível 3		Nível 4		Nível 5	
	%	(e.p.)*	%	(e.p.)*	%	(e.p.)*	%	(e.p.)*	%	(e.p.)*	%	(e.p.)*
Brasil_9+	10	(4)	30	(1)	35	(2)	19	(1)	5	(1)	1	(0.5)
Brasil_8	32	(2)	40	(1)	21	(2)	6	(1)	1	(0.2)	0	0
Brasil_7	57	(2)	32	(2)	10	(1)	1	(0.5)	0	0	0	0

TABELA 2- PROPORÇÃO DE DIVERSAS SÉRIES NOS NÍVEIS DE PROFICIÊNCIA.

FONTE: PISA 2000 – RELATÓRIO NACIONAL(2000)

[26] Ver os cinco níveis estabelecidos no anexo 3. Para maiores detalhes consultar o site do PISA: http://www.inep.gov.br/download/internacional/pisa/PISA2000.pdf

[27] Para maiores detalhes consultar o site do PISA abaixo:
http://www.inep.gov.br/download/internacional/pisa/PISA2000.pdf

Kato (2007), na apresentação de seu livro, afirma que os pesquisadores da área de leitura em LE da PUCSP constataram que no ensino de leitura instrumental, “muito das dificuldades dos aprendizes devia-se não ao desconhecimentos da língua estrangeira, mas principalmente à sua inabilidade de interagir com o texto escrito na própria língua materna.” .

Tendo em vista esse cenário, pode-se deduzir que, de maneira geral, o leitor real dessa pesquisa não corresponde ao leitor ideal caracterizado acima, no sentido de que o leitor real, provavelmente, não é capaz de selecionar e utilizar as funções mentais e os processos de leitura adequados e necessários no processamento textual, o que pode comprometer a construção de sentido do texto e, conseqüentemente, a realização eficaz da tarefa solicitada.

O fato de o leitor real não corresponder ao leitor ideal, conceituado antes, foi adotado como um pressuposto a ser considerado tanto na definição do construto, quanto na elaboração da escala de medida e dos descritores dos níveis da escala ordinal.[28] Na prática, isso significa que o ponto de partida para a elaboração dos descritores e da chave de respostas utilizada na correção foi o modelo de leitor ideal, mas que as expectativas em relação ao resultado final e os ajustes em termos de severidade de correção foram redefinidos e acordados verbalmente pelos corretores no treinamento, no sentido de se levar em consideração a presença factual do leitor real.

Com a identificação do leitor ao qual esta pesquisa se refere, faz-se necessário estabelecer que tipo de texto espera-se que este leitor compreenda.

2.3 TEXTO

O item anterior refere-se ao leitor, um dos três elementos envolvidos na compreensão de leitura, apresentados no item 2.1 acima. Outro desses três elementos é o texto, que constitui o objeto da compreensão que o leitor se propõe a fazer através da leitura. O termo texto, no contexto desta pesquisa é empregado, neste primeiro momento, como referência a todo texto escrito, mas

[28]Ver cap. 5, item 5.1.2 e 5.1.3.

na medida em que for sendo estabelecido o tipo de texto com o qual se pretende trabalhar, o termo também vai se especializando, reduzindo a abrangência do seu significado, até passar a designar especificamente o tipo de texto escolhido.

Em se tratando de textos, pode-se perceber uma variedade muito grande de formas, extensão, conteúdo, entre outras características. Há também, uma diversidade de classificações estabelecidas para agrupar os textos de acordo com as semelhanças de características que eles apresentam.

De acordo com Koch e Elias "no processo de leitura e construção de sentido dos textos, leva-se em conta que a escrita/fala baseiam-se em formas padrão e relativamente estáveis de estruturação". A lista é bastante extensa,

> tanto que estudiosos que objetivaram o levantamento e a classificação de gêneros textuais desistiram de fazê-lo, em parte porque os gêneros existem em grande quantidade, em parte porque os gêneros, como práticas sociocomunicativas, são dinâmicos e sofrem variações na sua constituição, que, em muitas ocasiões, resultam em outros gêneros, novos gêneros. (2006, p.101).

Da mesma forma, há também entre os autores diferenças de emprego de termos como *gênero de texto*, *tipo de texto, estrutura de texto.* De acordo com o PREL, por exemplo,

> *gênero de texto* pode ser classificado de muitas formas, tais como ficção, não ficção, contos de fada, fábulas, e peças. *Estrutura de texto* refere-se a padrões familiares que estabelecem as interrelações entre as idéias de um gênero. (2005, p. 10).

Apesar das diferenças de classificação existentes e da impossibilidade de se estabelecer uma única classificação como sendo a melhor de todas, é possível e necessário, nesta pesquisa, que se faça algum tipo de classificação em função das características presentes nos textos, uma vez que as características vão influenciar na forma de abordagem que o leitor vai fazer desses textos. A esse respeito Snow (2002, p. 14) afirma que "as características do texto têm grande efeito sobre a compreensão." Segundo ela, a compreensão não é simplesmente uma questão de extração de significado do texto, mas envolve a construção que o leitor faz de representações do texto, que incluem a estrutura superficial do texto (ou seja, as palavras exatas que o

texto contém), a estrutura básica do texto (isto é, as idéias representativas dos significados do texto) e ainda os modelos de representação mental que estão embutidos no texto. Portanto, segundo ela, características tais como o estilo empregado no discurso, o tipo de vocabulário, a estrutura lingüística, entre outras, podem facilitar ou dificultar a compreensão que um leitor tem de um texto.

Sendo assim, mais importantes que a nomenclatura dada na classificação do gênero de texto utilizado nesta pesquisa são as características que ele apresenta e a compreensão que se espera que o leitor tenha, em vista dos propósitos da avaliação. Entretanto, como se faz necessário determinar os conceitos teóricos que embasam as decisões tomadas, serão relacionados a seguir alguns conceitos de gênero, estabelecido o gênero a ser utilizado na elaboração do teste e apresentada a razão de sua escolha.

2.3.1 O Conceito de Gênero

De acordo com o dicionário “Webster Third”, *gênero* é “um tipo ou categoria distintiva de composição literária”. No entanto, este termo tem sido correntemente usado para designar categorias distintivas de qualquer tipo de discurso, seja ele falado ou escrito, esteja ele objetivando ou não estar inserido no campo literário. (SWALES, 1990, p. 33).

Esse autor discute em seu livro “Genre Analysis” (1990) o uso do termo *gênero* em quatro áreas específicas, dentre as diversas possíveis: folclore, estudos literários, lingüística e retórica. A discussão da adequação do uso do termo *gênero* por diferentes áreas de pesquisa não se faz relevante para o presente estudo, cabendo destacar que, as definições e usos do termo que interessam e são utilizadas nesta pesquisa provêm da lingüística.

Swales (1990, p.40) afirma que na lingüística a relação entre os conceitos de *gênero* e *registro* ainda não está muito clara. Segundo ele, *registro* é uma categoria contextual, isto é, é uma variação funcional da linguagem que correlaciona grupos de características lingüísticas com

características de situações que são recorrentes. Esta categoria é geralmente analisada por meio de três variáveis: “field” (literalmente = campo), que indica onde o discurso opera, seu conteúdo, idéias e foco; “tenor” (literalmente = teor), que indica a relação entre os sujeitos e “mode” (literalmente = meio), que indica o canal de comunicação (oral ou escrita).

De acordo com Swales (1990, p.40) os lingüistas, em geral, tendem a não abordar a questão da conceituação de gênero. No entanto, ele apresenta as afirmativas de Martin e Couture sobre assunto da seguinte forma:

a) segundo Martin (1985) os “*gêneros* são realizados através dos registros e registros, por sua vez, são realizados através da linguagem” (tradução nossa). Martin emprega o termo *gênero* para designar os diferentes tipos de atividades lingüisticamente realizadas em uma determinada cultura;
b) Couture (1986) entende que *registros* impõem limites no que concerne ao vocabulário e sintaxe da língua, e os limites impostos pelos *gêneros* acontecem no nível da estrutura do discurso. Os *gêneros* só podem ser percebidos em discursos completos ou que possam ser considerados como tal, diferentemente dos *registros*.

Embora considerando que a sua definição de *gênero* possa não ser totalmente adequada aos meios acadêmicos e de pesquisa, Swales a propõe da seguinte maneira:

> Um gênero consiste em uma classe de eventos comunicativos, cujos membros compartilham alguns grupos de propósitos comunicativos. Estes propósitos são reconhecidos pelos membros especialistas da comunidade de discurso matriz, e assim constituem a lógica para o gênero. Essa lógica molda a estrutura esquemática do discurso e influencia e restringe escolhas de conteúdo e estilo. Propósito comunicativo é tanto um critério restrito quanto um (critério) que funciona para manter o escopo de um gênero como concebido aqui, centrado especificamente na ação retórica comparável. Além do propósito, exemplares de um gênero exibem vários padrões de similaridade em termos de estrutura, estilo, conteúdo, e público alvo. Se todas as expectativas de alta probabilidade são concretizadas, o exemplar será visto como prototípico pela comunidade de discurso matriz. Os nomes dos gêneros herdados e produzidos pelas comunidades de discurso e importados por outros constituem

> comunicações etnográficas valiosas, mas geralmente precisam de mais validação."[29] (1990-p. 58, tradução e grifo nosso).

Mesmo entendendo que a terminologia empregada para cada gênero ainda precise de mais validação por parte da comunidade discursiva, conforme afirma Swales, acima, esta pesquisa adota a terminologia estabelecida por Goldman e Bisanz, na medida em que parece ser a mais adequada neste momento, na opinião da autora da pesquisa.

Para estabelecer sua classificação de textos científicos Goldman e Bisanz (2002, p. 21- 26) consideram três papéis, ou funções, socioculturais principais das comunicações científicas: a "comunicação entre cientistas, a popularização da informação científica para aqueles que estão fora da comunidade científica, e educação científica formal" (tradução nossa). Entre estas três funções, apenas as duas primeiras interessam a esta pesquisa.

Na comunicação entre cientistas elas distinguem dois grandes grupos de gênero: o *formativo (formative)* e o *integrativo (integrative)*. Os primeiros "documentam e moldam o pensamento dos cientistas e refletem a vanguarda no campo científico"; o segundo "são sínteses do que é amplamente conhecido e aceito sobre uma área temática particular". Dentro do gênero formativo, interessa a esta pesquisa os textos que as autoras chamam de "artigos de periódicos referendados, incluindo relatórios de pesquisa empírica, revisão crítica de uma área temática, e formulações teóricas" (*Refereed journal articles,*

---

[29] "A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. Communicative purpose is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content, and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community. The genre names inherited and produced by discourse communities and imported by others constitute valuable ethnographic communications, but typically need further validation."

including reports of empirical research, critical reviews of a topic área, and theoretical formulations). (traduções nossa).

Na função de popularização de informações científicas, as autoras também dividem os gêneros em dois grupos: textos voltados para o conhecimento público (public awareness), cuja intenção é aumentar a conscientização do público sobre as informações científicas, e textos voltados para o entendimento do público e aprendizagem informal (public understanding and informal learning), que busca melhorar o entendimento que o público tem a respeito das informações científicas. Neste segundo gênero, interessam os artigos (feature articles) e resumos de relatórios/comentários críticos (summary reports/critical commentaries)[30]. (traduções nossas).

A escolha dos gêneros de texto acima foi feita levando-se em consideração o público alvo da pesquisa, isto é, candidatos a mestrado ou doutorado, os quais, a despeito da área de pesquisa, terão que utilizar textos científicos recorrentes e imprescindíveis no embasamento teórico de um grande número de pesquisas, seja em ciências tecnológicas, biológicas ou humanas. Os gêneros apresentados acima são os que, no entendimento desta autora, representam melhor os textos gerados pela e para a academia, contexto sociocultural em que está inserida esta pesquisa.

Graesser, León e Otero (2002, p. 4), entre as várias razões para o estudo de textos científicos, apresentam algumas de especial importância, que também justificam o seu uso nas avaliações propostas. Segundo eles, o conhecimento do conteúdo de textos científicos é útil para o leitor e a promoção de educação científica é do interesse, em princípio, de todos os países e culturas, uma vez que este conhecimento científico pode tornar os cidadãos mais esclarecidos sobre questões que podem proporcionar melhores condições de vida e saúde a todos eles. A afirmativa de que o conteúdo de textos científicos é importante para o leitor é especialmente verdadeira em se tratando do meio acadêmico. Além disso, segundo os autores acima, “este gênero de texto tem uma forma característica de organizar e explicar o material. Freqüentemente assume-se que coerência e compreensão estão intimamente

[30] Para a lista completa de gêneros apresentada por Goldman e Bisanz, consultar a obra indicada na bibliografia (p. 25-6).

relacionadas" (tradução nossa). E ainda que afirmem que textos científicos são difíceis de entender em profundidade, a necessidade da educação científica e a necessidade de se trabalhar com esse gênero de texto na academia o tornam particularmente apropriado para avaliações, em que se pretende medir, também, a capacidade de utilizá-los.

Para finalizar, é importante ressaltar que todos os conceitos apresentados neste capítulo sobre gênero de texto, leitor e o que seja compreensão de leitura são, de alguma forma, relevantes para esta pesquisa. Nela considera-se que nos processos de leitura compreensiva deve haver a utilização de processos cognitivos e metacognitivos por parte do leitor, tanto na leitura do texto quanto na realização das tarefas propostas para a leitura, levando sempre em consideração o contexto em que a leitura acontece. Isto se dá porque parte-se do princípio de que os leitores alvo da pesquisa são o que, segundo o PREL podem ser considerados bons leitores, por serem capazes de ativar processos metacognitivos durante a leitura. Como se pode perceber ao longo do capítulo, a figura apresentada no item 2.1, proposta pelo RAND, representa as dimensões que estão envolvidas na avaliação de leitura proposta neste estudo, com a diferença de que, exatamente por tratar-se de um processo de avaliação, os limites do contexto sociocultural são determinados, não pelo leitor, mas por um órgão educacional responsável, pela instituição em que esta avaliação é desenvolvida e aplicada e, logicamente, pelo(s) elaborador(es) da avaliação. Sendo assim, considerando que o assunto desta pesquisa não é apenas a leitura em si, mas também as avaliações de leitura em língua inglesa é importante que, a seguir, alguns conceitos relacionados às avaliações sejam também apresentados.

3 REVISÃO BIBLIOGRÁFICA DE AVALIAÇÃO

Esta pesquisa, além de ser uma pesquisa sobre leitura, é mais precisamente uma pesquisa sobre **avaliações** de leitura. Por isso, assim como houve a necessidade de apresentar uma revisão bibliográfica sobre a leitura, é preciso também que se apresente uma revisão bibliográfica dos conceitos sobre avaliação, que são relevantes para esta pesquisa e que servem de base para as escolhas feitas e para as decisões tomadas na elaboração, correção e análise dos testes. É exatamente isso que será feito neste capítulo.

Entretanto, é importante salientar que alguns dos conceitos sobre avaliação só serão apresentados em outros capítulos, na medida em que forem necessários para a compreensão do assunto tratado. Isto foi feito para evitar que o leitor desta pesquisa precisasse voltar ao presente capítulo para entender determinados assuntos que estão diretamente relacionados àqueles conceitos mais específicos.

3.1 MEDIÇÃO – AVALIAÇÃO – TESTE

Embora os termos avaliação, teste e medição sejam constantemente utilizados no meio educacional, nem sempre as nuanças de definição estão claras para aqueles que as utilizam. Como esta é uma pesquisa que trata especificamente deste assunto, é importante esclarecer a definição aplicada a cada um deles aqui.

Segundo Bachman (1991, pp. 18-24):

Mensuração (Measurement)[31]: “Mensuração nas ciências sociais é o processo de quantificar as características de pessoas de acordo com

[31] Mensuração não é um termo muito freqüente na área de educação em português. No entanto, como é um termo bastante corrente no inglês, para traduzir “measurement”, a autora optou por “mensuração”, por ser a palavra cujo significado mais se aproxima do sentido que o

procedimentos e regras explícitas. Esta definição inclui três aspectos distintos: quantificação, características, e procedimentos e regras explícitas."[32] (tradução da nossa).

Mensuração é geralmente baseada na observação naturalista do comportamento em um dado período de tempo, mas não vai necessariamente levar a resultados que comprovem as habilidades ou atributos específicos do comportamento do indivíduo. Por exemplo, alguém pode mensurar a habilidade oral em inglês de um indivíduo com quem conversa regularmente há muito tempo nesta língua e até mesmo graduar a sua produção oral dentro de alguma escala, de acordo com critérios de nivelamento pertinentes ao caso, no entanto, essa mensuração não poderá ser usada como prova dessa habilidade, uma vez que o avaliador não seguiu os procedimentos padrões exigidos por um teste de nivelamento. Devido a esse tipo de limitação que a mensuração apresenta, os testes de linguagem são considerados mais apropriados para garantir que a amostra de linguagem obtida de fato serve para os propósitos pretendidos.

Avaliação (Evaluation): "Avaliação pode ser definida como a coleta sistemática de informações para o propósito de tomada de decisões."[33] (WEISS[34], 1972, apud BACHMAN, 1991, p.22, tradução nossa).

O termo "assessment" é empregado como sinônimo de "evaluation" e o termo correspondente em português também é "avaliação". Uma das características da avaliação é a coleta de informações confiáveis e relevantes, sendo que as informações não precisam, necessariamente ser quantitativas, assim como não abrangem todos os tipos de testes, que podem ser usados com propósitos descritivos, por exemplo.

---

termo em inglês carrega e por não haver outro termo que, além de definir melhor, seja diferente do usado para a tradução de "evaluation" (avaliação).

[32] "Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules. This definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures."

[33] "Evaluation can be defined as the systematic gathering of information for the purpose of making decisions."

[34] WEISS, C. H. **Evaluation Research: Methods for Assessing Program Effectiveness**. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Teste (Test): “um teste é um instrumento de medição elaborado para obter uma amostra do comportamento de um indivíduo. Como um tipo de medição, um teste necessariamente quantifica características de indivíduos de acordo com procedimentos explícitos.”[35] (BACHMAN, 1991, p.20, tradução nossa).

A grande vantagem do teste reside na sua capacidade de obtenção de tipos de comportamento possíveis de serem interpretados como sendo evidências da habilidade testada. Por seguirem procedimentos pré-estabelecidos e serem preparados especialmente para o propósito utilizado é que os testes de linguagem podem ser considerados confiáveis, significativos e úteis.

Segundo Brown, teste

> é um método para medir a habilidade ou o conhecimento de uma pessoa em um determinado assunto. [...] É um conjunto de técnicas, procedimentos e itens que constituem um instrumento,” (2001, p.384, tradução nossa)

que exige algum tipo de desempenho ou atividade por parte do aluno. E é este desempenho ou atividade que será avaliado. Esta definição proposta por Brown é abrangente e clara, mas nem por isso simples de se realizar.

É possível estabelecer uma relação entre estes três termos com as seguintes proposições:

a) nem toda avaliação é uma mensuração ou um teste;
b) nem toda mensuração é uma avaliação ou um teste;
c) nem todo teste é uma avaliação, e;
d) todo teste é necessariamente uma mensuração.

O esquema proposto por Bachman (1991, p. 23, fig. 2.1) ajuda a entender melhor as proposições acima.

---

[35] “a test is a measurement instrument designed to elicit a specific sample of an individual’s behavior. As one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures.”

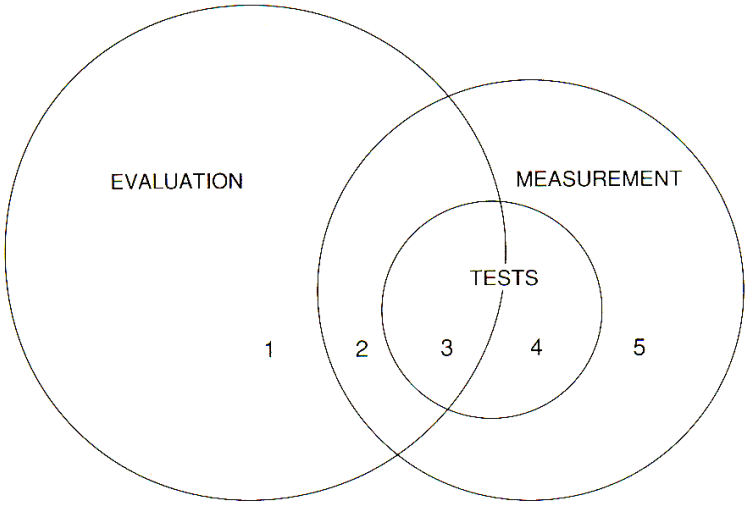


Figure 2.1 Relationships among measurement, tests, and evaluation

FIGURA 2 – RELATIONSHIPS AMONG MEASUREMENT, TESTS, AND EVALUATION.
FONTE: BACHMAN (1991)

Os testes de que trata esta pesquisa correspondem, no esquema de Bachman, ao tipo 3, ou seja, os testes desta pesquisa são necessariamente avaliações, porque a coleta de informações é sistemática e se destina a decidir se o desempenho do candidato pode ser considerado suficiente ou insuficiente, dentro dos limites pré-estabelecidos como tal.

3.2 O PROCESSO DE AVALIAÇÃO

Em se tratando de avaliações, mais especificamente de testes, um dos pontos mais importantes é assegurar que os testes tenham realmente a qualidade esperada, isto é, que testem aquilo a que se propõem testar da melhor forma possível, e que reflitam, de fato, o conhecimento dos indivíduos que estão sendo avaliados nos resultados obtidos. Para que isso aconteça, existem certos critérios a serem observados na elaboração dos testes, que garantem que os mesmos vão avaliar o domínio pretendido, sem incluir fatores desnecessários ou indesejados, ou permitir que outros fatores possam interferir nos resultados e/ou na interpretação dos mesmos. Diferentes autores apontam um número diferente de critérios, ou qualidades, como sendo mais relevantes

na qualificação dos testes. Entre eles, dois critérios são recorrentes na literatura que trata do assunto: a confiabilidade e a validade. Esses critérios envolvem questões relativas ao processo de elaboração, aplicação e correção de testes, tais como a adequação dos testes elaborados quanto aos objetivos institucionais em que os testes são aplicados, quanto ao seu construto e quanto à medida utilizada na pontuação e análise dos resultados dos testes. Sendo assim, estes dois critérios e as questões relativas a cada um deles, pertinentes a este trabalho, são tratados a seguir.

3.2.1 Validade

Embora a questão da validade esteja sempre em pauta nos estudos sobre testes e, implícita ou explicitamente, esteja inserida na atividade de elaboração de testes em si, nem sempre ela é considerada ou conscientemente verificada pelos elaboradores de teste em geral.

"O propósito da validação no teste de língua é assegurar a defensibilidade e justiça das interpretações baseadas no desempenho do teste." (MACNAMARA, 2000, p.48, tradução nossa). A responsabilidade que envolve a elaboração e aplicação de testes é muito grande. No caso de testes de suficiência, o propósito é medir um conhecimento considerado necessário a quem pretende o grau de mestre ou doutor e os resultados são utilizados para decidir quem tem ou não conhecimento suficiente para ler, compreender e utilizar as (ou parte das) informações encontradas nas leituras em LE, que forem pertinentes a cada caso, na elaboração do texto (dissertação ou tese) referente a cada pesquisa realizada. Em alguns casos, esses testes são eliminatórios no processo de ingresso no curso, em outros são condição *sine qua non* na obtenção do grau de mestre ou doutor. Faz-se, portanto, necessário que os resultados obtidos sejam considerados justos, caso contrário, todo processo é passível de ser contestado e invalidado.

De acordo com McNamara,

> A pesquisa conduzida para validar procedimentos de testes pode acompanhar o desenvolvimento do teste, e é freqüentemente feito pelos próprios elaboradores de teste; isto é, ele pode começar antes do teste se tornar operacional. Validação idealmente continua através a vida do teste, conforme novas questões a respeito da sua validade são levantadas, normalmente no contexto da pesquisa de teste de língua.[36] (2000, p.49)

Segundo Douglas Brown, na avaliação de adequação de um teste, "De longe, o critério mais complexo de um bom teste é a **validade**, o quanto o teste de fato mede o que se pretende medir."[37] (2001, p.387, tradução nossa). E pelo grande número de estudos que se encontra sobre esse critério, embora aqui estejam citados apenas alguns deles, pode-se perceber que ele não é o único a ter esse entendimento sobre o assunto.

Bachman (1991, p. 24-25) afirma que "Se nós temos que interpretar a pontuação num dado teste como um indicador da habilidade de um indivíduo, aquela pontuação deve ser tanto confiável quanto válida."38 Mas que, "A qualidade mais importante da interpretação ou uso do teste é a validade". Ele considera que "enquanto a confiabilidade é uma qualidade dos escores do teste em si, validade é uma qualidade da interpretação e uso do teste." (tradução e grifo nosso). Ou seja, espera-se que os resultados dos testes sejam confiáveis e que as interpretações e os usos que serão feitos das respostas sejam válidos. Mas, para que as interpretações sejam válidas, os resultados devem ser confiáveis. Afirmar isso equivale a que dizer que, em certa medida, a validade depende da confiabilidade, como afirma Davis et al no Dicionário de teste de linguagem. De acordo com este dicionário, validade é:

> a qualidade que mais afeta o valor de um teste, anterior a, embora **dependente da**, confiabilidade. Uma medida é válida se faz o que ela tem a intenção de fazer, que é tipicamente agir como um indicador de

---

[36] "The research carried out to validate test procedures can accompany test development, and is often done by the test developers themselves; that is, it can begin before the test becomes operational. Validation ideally continues through the life of the test, as new questions about its validity arise, usually in the context of language testing research."

[37] "By far the most complex criterion of a good text is **validity**, the degree to which the test actually measures what is intended to measure."

[38] "If we are to interpret the score on a given test as an indicator of an individual's ability, that score must be both reliable and valid."

> um conceito abstrato (por exemplo, altura, peso, tempo, etc.) que ela afirma medir. A validade de um teste de linguagem portanto é estabelecida pelo grau com o qual ele é bem sucedido em provar uma representação concreta precisa de um conceito abstrato (por exemplo proficiência, realização, aptidão).Os tipos de validade referidos mais comuns são[39]: conteúdo; construto; concorrente; preditiva."[40] (DICTIONARY OF LANGUAGE TESTING, 2002, p.221, tradução e grifo da nosso)

Segundo Alderson, Clapham e Wall (1995, p. 171), "validade pode ser estabelecida de várias maneiras diferentes, que leva a maioria dos escritores neste tópico a falar de diferentes tipos de validade." Pasquali (2003, p. 159) apresenta três parâmetros de validade como os mais conhecidos atualmente: validade de construto, de conteúdo e de critério.

No que se refere à validade de construto, que será a primeira das três a ser vista, uma das formas utilizadas para se assegurar que um teste seja justo é verificar se o uso que se faz dos testes está de acordo, ou se é adequado, entre outras coisas, quanto aos objetivos traçados pela instituição que os utiliza. Ou seja, um teste para ser válido precisa estar de acordo com os propósitos estabelecidos pela instituição para a / na qual ele foi desenvolvido.

3.2.1.1 Adequação do Teste Quanto Aos Objetivos Institucionais

Em contextos educacionais, o objetivo principal da leitura é construir uma representação mental que possa ser recuperada e utilizada quando a informação contida no texto for, de alguma forma, necessária no futuro. Essa

---

[39]Esses quatro tipos de validade são também os propostos por CROMBACH e MEEHL (1955) em seu artigo *Construct validity in psychological tests.* As validades **concorrente** e **preditiva** são agrupadas por eles sob a denominação de validade de critério-orientado (criterion-oriented) e simplesmente validade de **critério** por PASQUALI (2003, cap. 6).

[40] "The quality which most affects the value of a test, prior to, though dependent on, **reliability**. A measure is valid if it does what it is intended to do, which is typically to act as an indicator of an abstract concept (for example height, weight, time, etc.) which it claims to measure. The validity of a language test therefore is established by the extent to which it succeeds in providing an accurate concrete representation of an abstract concept (for example **proficiency, achievement, aptitude**).The most commonly referred to types of validity are: **content; construct; concurrent; predictive**."

capacidade de construir uma representação mental do que se lê é avaliada de diversas maneiras, dependendo do objeto de estudo e do nível de capacidade de compreensão exigido em cada caso.

Para que um processo de avaliação exista é necessário haver primeiramente um motivo para se avaliar. Bachman e Palmer (1996 –p.95) afirmam que o "uso primário dos testes de linguagem é para fazer inferências sobre a capacidade de linguagem". Seguindo esta mesma linha, vários autores concordam com o fato de que o propósito básico dos testes é tomar decisões com base nas inferências feitas. No presente estudo este propósito básico também deve existir. Assim sendo, deve-se perguntar qual o objetivo de um teste e como ele se situa na classificação dos autores.

De acordo com Mcnamara (2000, p. 05) os testes podem diferir na maneira como são elaborados e na razão para a qual são utilizados, ou seja, eles podem diferir quanto ao método e quanto ao propósito. Em relação ao método, segundo ele, os testes podem ser divididos, de forma bem generalizada, entre o que ele chama de *paper-and-pencil tests* (literalmente = testes de papel e lápis) e *performance tests* (testes de desempenho).

Os primeiros, também denominados *testes indiretos,* são testes de linguagem tradicionais, escritos, geralmente usados para avaliar os componentes da língua, tais como gramática ou vocabulário, separadamente, ou para avaliar as habilidades receptivas da língua (audição e leitura) e, de acordo com o *Dictionaty of Language Testing* (2002, p. 81) são testes "que **não requerem** que o examinando realize tarefas que refletem diretamente o tipo de uso de linguagem que é o alvo da avaliação". (tradução e grifo nosso).

Os *testes de desempenho*, ou também chamados *testes diretos*, de acordo com McNamara (2000, p. 06), são testes em que as "habilidades de linguagem são avaliadas em um ato de comunicação" e "são mais comumente testes de oralidade e escrita, em que uma amostra mais ou menos extensa é obtida do examinando, e julgada por um ou mais corretores treinados usando um procedimento de correção acordado." Esses testes são definidos pelo *Dictionaty of Language Testing* (2002, p. 144) como "um teste em que a habilidade dos candidatos para realizar determinadas tarefas, normalmente associadas com necessidades de trabalho ou estudo, é avaliada". De acordo

com este dicionário a aplicação desse tipo de teste só é possível “onde há uma clientela relativamente homogênea com conhecida e relativamente específica necessidade de uso da linguagem." (traduções nossa).

A classificação dos possíveis propósitos varia um pouco de acordo com cada autor. McNamara (2000,p.06) divide os propósitos dos testes em dois grandes grupos: *resultados alcançados*, que são associados ao processo de instrução e, portanto dizem respeito ao que foi aprendido no período anterior ao teste (passado) e *proficiência*, que é voltado para o futuro uso da linguagem, sem ter, necessariamente, relação com algum processo de aprendizagem.

Bachman (1990, p. 58-61) propõem uma classificação com cinco tipos de propósitos, sendo que os dois últimos estão reunidos em um e todos eles podem ser relacionados a um ou ambos os grupos propostos por McNamara. São eles: *seleção*, usado para determinar quem deve ser admitido em um programa educacional ou em um emprego (proficiência); *nivelamento*, usado para determinar em que nível de instrução um indivíduo deve ser colocado (proficiência); *diagnóstico*, usado para identificar e dar informações sobre os pontos fortes e fracos de um aluno em áreas específicas, com o objetivo de direcionar os estudos do aluno para melhorar seus pontos fracos (proficiência ou resultados alcançados); *progresso e dar nota/classificação*, usado para dar informação sobre o progresso do aluno e, conseqüentemente, sobre o professor e o método de ensino utilizado, informação esta, normalmente, traduzida em termos de escores dos testes (resultados alcançados).

A classificação de Alderson (1995, p.11 - 12) também divide os testes em cinco grandes categorias, assim como o faz Bachman, por outro lado, essas categorias não se encaixam nos dois grupos de McNamara, mas as incorporam como parte das cinco, que são: *nivelamento*, elaboradas para acessar o nível de capacidade de linguagem de um aluno, para que se possa colocá-lo no curso ou nível apropriado; *progresso*, usadas em diferentes estágios de um curso para saber o que os estudantes aprenderam; *resultados alcançados*, semelhantes aos testes da categoria de progresso, apresentam a diferença de, em geral, serem aplicados no final do curso; *proficiência*, elaboradas para testar a capacidade dos alunos com diferentes contextos de aprendizagem de linguagem e *diagnósticos*, que visam identificar as áreas em

que os alunos precisam de atenção extra, quer sejam áreas gerais (e.g., estruturas de sentenças), ou específicas (e.g., posição dos adjetivos na sentença).

Considerando a classificação de testes apresentada acima, pode-se dizer que os objetivos institucionais que embasam o presente trabalho são classificados quanto ao método como um teste de desempenho, porque, apesar de McNamara apresentá-lo como um tipo mais comumente usado em oralidade e escrita, esta autora considera que este é um teste utilizado em uma situação de uso comunicativo da linguagem, cujo foco é uma habilidade receptiva; em que a clientela pode ser considerada homogênea; com uma necessidade de uso da linguagem específica e conhecida e um teste no qual o desempenho será julgado por corretores treinados e em acordo. Quanto ao propósito, o teste é classificado de acordo com McNamara e Bachman, pelo fato de suas classificações parecerem mais adequadas aos propósitos em questão. Sendo assim, de acordo com McNamara a avaliação é de proficiência, por estar interessada no uso futuro da linguagem, sem se preocupar com a instrução pela qual o candidato passou para chegar ao seu nível de conhecimento no momento do teste. Mais especificamente, segundo Bachman e Palmer, esta é uma avaliação de *seleção*, pois visa determinar quais candidatos têm o conhecimento de língua desejado para o cumprimento do programa de pós-graduação estabelecido pela instituição.

O ponto de partida é saber a real necessidade de uma avaliação de proficiência para a seleção de candidatos às vagas dos programas de mestrado e doutorado de uma universidade. Procura-se aqui analisar esta necessidade sob dois pontos de vista, que se têm como fundamentais para qualquer instituição: o prático e o legal. Do ponto de vista prático, de maneira geral, fatores como a falta de literatura técnica produzida na ou traduzida para a língua materna e a necessidade de busca de informação nas mais diversas fontes fazem com que a capacidade de leitura em uma língua estrangeira seja fundamental para os candidatos a esses programas de pós-graduação, independentemente da área a ser pesquisada. Sendo assim, existe uma necessidade de ordem prática de obtenção de determinadas informações por meio de uma língua estrangeira. Conseqüentemente, faz-se necessário

verificar se o candidato é capaz de compreender textos escritos que só estejam disponíveis em outro idioma, e se esta capacidade é suficiente para que ele possa selecionar e utilizar de forma adequada as informações que sejam relevantes para atingir seus objetivos acadêmicos, contidas nesses textos.

Sob o ponto de vista legal, o Regimento Geral da UFPR, por exemplo, afirma no Art. 52. e Art. 53 – ambos no item IV, que para obtenção do grau de mestre e doutor, respectivamente, o regimento estabelecerá, entre outras condições a exigência de que o candidato seja aprovado em uma prova de capacidade de tradução[41] de um texto específico em língua estrangeira.

Na seleção dos candidatos aos cursos de pós-graduação *strictu sensu* da Universidade Federal do Paraná, a Resolução do CEPE n.º 62/03 diz, no Art. 34 item (a), que para admissão e/ou obtenção do título pretendido, o candidato deverá satisfazer as exigências de "ser aprovado em processo seletivo instituído pelo colegiado do programa e em teste de suficiência em língua estrangeira moderna, em uma ou duas línguas estrangeiras modernas;". Diz ainda no §2º, que "para efeito desta Resolução estende-se por teste de suficiência em língua o que se realiza com o objetivo específico de verificar se o aluno é suficiente em leitura compreensiva de textos de divulgação científica ou retirados de revistas científicas."

Para que os testes de suficiência de fato verifiquem o que é esperado conforme a resolução do CEPE, acima, é fundamental que ele possua as qualidades anteriormente mencionadas de confiabilidade e de validade, entre as quais está a validade de construto, da qual se falará a seguir.

---

[41] Embora o termo *tradução* não esteja definido na Resolução, para efeitos desta pesquisa, entende-se aqui este termo como sendo o resultado de uma leitura compreensiva de um texto, que é expresso da maneira mais fiel possível ao texto em LE, em oposição à possível decodificação do código lingüístico, que envolve muito mais um trabalho de pesquisa em dicionário do que o entendimento e a construção de sentido a partir do texto lido.

3.2.2 Validade de construto

Além da importância de se validar um teste pelo uso que se faz dele, é preciso considerar que existem outras variáveis que influenciam direta ou indiretamente na elaboração de uma avaliação, qualquer que seja, e que elas são bastante complexas. Essas variáveis são chamadas de *construto*. Este construto varia de acordo com o que se pretende testar e deve ser um indicador da habilidade que se pretende medir.

Bachman e Palmer afirmam que

> Para justificar uma interpretação de um escore em particular, nós precisamos fornecer evidências de que o escore do teste reflete a(s) área(s) da habilidade de linguagem que nós queremos medir, e muito pouco além disso. [...] O termo *validade de construto* é portanto usado para se referir-se à até que ponto nós podemos interpretar um determinado escore de teste como um indicador da(s) habilidade(s), ou construto(s), que nós queremos medir."[42] (1996, p. 21, tradução nossa).

Deste modo, na questão da adequação do teste quanto ao construto deve-se considerar se o que se pretende testar está realmente de acordo com o que diz a teoria da compreensão de leitura que embasa a elaboração desse teste. É imprescindível saber exatamente o quê se pretende medir através desse teste, por exemplo, se a capacidade de entender a idéia geral de um texto, de encontrar informações específicas, ou, talvez, a capacidade de identificar as opiniões do autor do texto através das idéias expostas. E, para que se possa definir a natureza do construto é preciso definir o próprio termo e saber o que significa dizer que um construto é válido.

Conforme o "Dictionary of language testing", construto é

---

[42] "In order to justify a particular score interpretation, we need to provide evidence that the test score reflects the area(s) of language ability we want to measure, and very little else. […] The term *construct validity* is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure."

> o traço ou traços[43] que um teste pretende medir. Um construto pode ser definido como uma capacidade ou um conjunto de capacidades que estarão refletidas no desempenho de um teste, e sobre as quais poderão ser feitas inferências com base na pontuação do teste. Um construto é geralmente definido em termos de uma teoria; no caso da linguagem, uma teoria da linguagem. Um teste, então, representa uma operacionalização da teoria. Validação do construto envolve uma investigação do que um teste na verdade mede e tenta explicar o construto. (1999, p.31, tradução nossa).[44]

Segundo Pasquali (2003, p. 159) a validade de construto pode também ser chamada de validade de conceito e "é considerada a forma mais fundamental de validade dos instrumentos psicológicos" de avaliação, porque "constitui a maneira direta de verificar a hipótese da legitimidade da representação comportamental dos traços latentes".

Brown (2001, p.389) sugere que uma forma de se verificar a validade de construto de um teste é perguntar se esse teste utiliza o construto teórico da forma como ele foi definido para o teste em questão. Ele afirma que, em princípio, qualquer assunto relacionado ao ensino e aprendizagem de linguagem envolve construtos teóricos e apresenta como exemplo de construtos a proficiência, a competência comunicativa e a auto-estima.

Alderson et al. (1995, p.17) afirmam que cada teoria contém construtos, ou seja, conceitos psicológicos, que são seus componentes principais e que derivam da teoria da habilidade a ser testada. Além disso, Alderson explica que construtos

> vêm de uma teoria de leitura, e eles são concebidos através dos textos que nós selecionamos, das tarefas que nós requisitamos aos leitores desempenhar, dos entendimentos que eles demonstram e das inferências que nós fazemos daqueles entendimentos,

---

[43] Traços ou traços latentes são usados nesta pesquisa com o significado de: uma variável não observável; um processo psicológico; atributos que não estão sujeitos à observação empírica e que, por isso, precisa ser observado por meio dos comportamentos que são a eles atribuídos, para que possam ser 'objetos' de uma abordagem científica.

[44] "The **trait** or traits that a test is intended to measure. A construct can be defined as an **ability** or set of abilities that will be reflected in test **performance**, and about which inferences can be made on the basis of test **scores**. A construct is generally defined in terms of a theory; in the case of language, a theory of language. A test, then, represents an operationalisation of the theory. Construct **validation** involves an investigation of what a test actually measures and attempts to explain the construct."

> tipicamente como refletidos nos escores."[45] (2002, p.117, tradução nossa).

Segundo ele, é importante entender que os construtos são abstrações definidas para determinados propósitos e não o que ele chama de "entidades psicologicamente reais", que existem, já definidas, em nossas mentes.

Ebel e Friesbie (1991) definem construto da seguinte forma:

> O termo *construto* refere-se a um construto psicológico, uma conceituação teórica sobre um aspecto do comportamento humano que não pode ser medido ou observado diretamente. Exemplos de construtos são inteligência, motivação pelo êxito, ansiedade, conquista, atitude, domínio e compreensão de leitura. Validação de construto é o processo de reunir evidência para respaldar a afirmação de que um determinado teste realmente mede o construto psicológico que os elaboradores pretendem que ele meça. O objetivo é determinar o significado dos escores do teste, para assegurar que os escores signifiquem o que esperamos/ supomos que eles devam significar." [46] (apud ALDERSON, 1995, p. 183, tradução nossa).

A respeito da definição do construto em uma medição, Wright diz que:

> A idéia de uma medida requer uma idéia de uma variável onde a medida é localizada. Se a variável é visualizada como uma linha, então a medida pode ser retratada como um ponto naquela linha. Essa relação entre uma medida e sua variável é ilustrada na figura.[47] (1979, p.1).

---

[45] "…come from a theory of reading, and they are realized through the texts we select, the tasks we require readers to perform, the understandings they exhibit and the inferences we make from those understandings, typically as reflected in scores."

[46] "The term *construct* refers to a psychological construct, a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly. Examples of constructs are intelligence, achievement motivation, anxiety, achievement, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean."

[47] "The idea of a measure requires an idea of a variable on which the measure is located. If the variable is visualized as a line, then the measure can be pictured as a point on that line. This relationship between a measure and its variable is pictured in Figure."

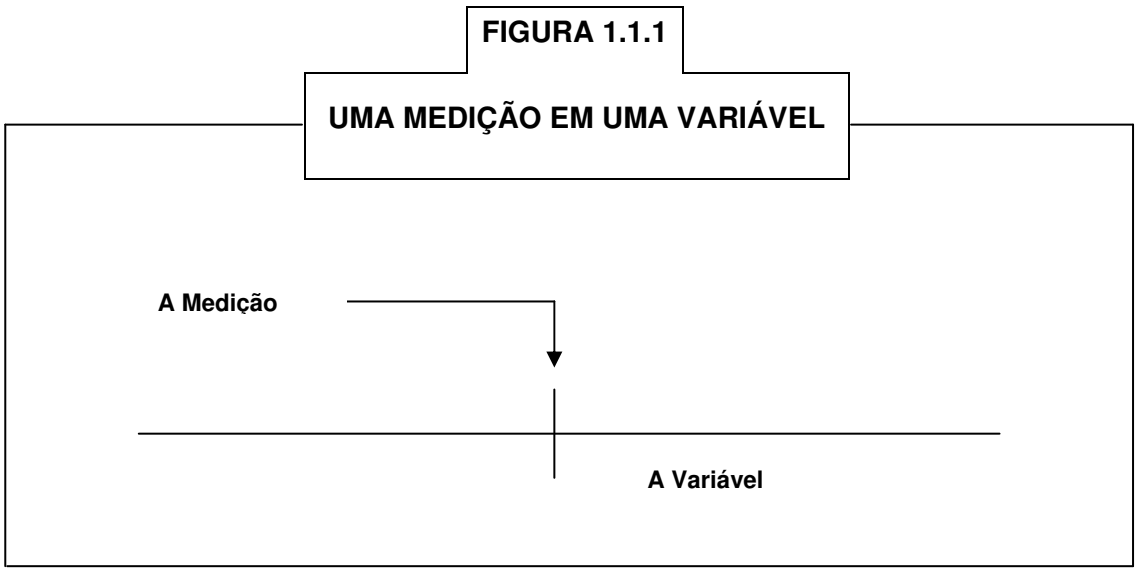


FIGURA 2: A MEASURE ON A VARIABLE
FONTE: "BEST TEST DESIGN" (1979)[48]

Em outras palavras, a linha que Wright e Stone (1979) estabelecem como sendo a variável é traçada a partir de uma idéia geral do que se quer medir. A variável que se quer medir pode ser, por exemplo, a capacidade de compreensão de leitura[49].

Um dos motivos pelos quais a definição da natureza do construto é importante em um teste é que construtos inadequados podem gerar inferências inadequadas, com a inclusão de um ou mais traços irrelevantes e/ou a exclusão de outros fundamentais para os propósitos da análise, por exemplo.

Bachman e Palmer (1996, p.117) apontam três propósitos básicos para a definição de construtos, quais sejam: servir de base para o uso do escore do teste, de acordo com o propósito para o qual o teste foi desenvolvido e aplicado; servir de guia para o desenvolvimento do próprio teste; e permitir que o elaborador e o usuário do teste demonstrem a validade do construto e das interpretações dos resultados do teste que forem feitas. Eles afirmam que a maneira pela qual se define o construto a ser medido vai ser determinada pelo tipo de inferência que se pretende fazer com base nos escores do teste.

[48] Ver figura original reproduzida no anexo 4.

[49] Isso diz respeito ao construto envolvido no presente estudo e que será abordado na seção 3.2.1.

Inferências essas, que dependem diretamente dos objetivos institucionais, tratados acima. Segundo eles:

> Uma consideração, portanto, é decidir quais componentes específicos da capacidade de linguagem serão incluídos na definição do construto. Em muitas situações de teste o usuário do teste vai querer fazer inferências sobre componentes específicos da capacidade de linguagem, e pode, portanto, definir o construto em termos daqueles componentes"[50] (1996, p.117, tradução nossa).

assim como ocorre em ambientes instrucionais, por exemplo. E neste caso, o elaborador do teste vai usar como base um programa instrucional específico. Porém,

> Em outros casos, como o uso de testes para determinar a admissão em um programa acadêmico, ou para tomar decisões sobre emprego, em que pode não haver um currículo instrucional, o elaborador do teste provavelmente irá basear a definição do construto nos componentes descritos em uma teoria de capacidade de linguagem.[51] (p.117, tradução nossa).

Bachman afirma que

> características físicas como altura, peso e cor dos olhos podem ser experienciadas diretamente através dos sentidos, e pode, portanto, ser definida por comparação direta com um padrão diretamente observável. (1990, p.41).

Por exemplo, pode-se sentir o peso de um objeto em uma medição de peso, ou a temperatura da água em uma medição de temperatura e pode-se comparar com outro peso e outra temperatura previamente estabelecida como padrão. No entanto, quando se trata de proficiência lingüística, não é possível experienciar a capacidade lingüística de outra pessoa; tudo o que se consegue

---

[50] "One consideration is thus to decide which specific components of language ability are to be included in the construct definition. In many testing situations the test user will want to make inferences about specific components of language ability, and may thus define the construct in terms of those components."

[51] "In other cases, such as the use of tests for determining admission into an academic program, or for making decisions about employment, where there may not be an instructional syllabus, the test developer will most likely base the definition of the construct on the components described in a theory of language ability."

é fazer inferências sobre essa capacidade, observando o comportamento que se presume ser influenciado por ela. Por exemplo, pode-se observar um comportamento influenciado pela capacidade de leitura de um aluno para *inferir* a sua capacidade de compreensão de texto.

Quando testa-se alguém, a intenção é estimar o conhecimento dessa pessoa em relação ao assunto testado. Em outras palavras, Wright e Stone (1979, p.1) afirmam que "nosso propósito é estimar sua localização na linha implícita pelo teste", que vai da capacidade mais baixa para a capacidade mais alta de compreensão.

De acordo com Wright e Stone (1979) :

> Para que um teste defina uma variável de capacidade mental, os itens dos quais o teste é constituído devem compartilhar uma linha de investigação. Esta linha comum e seu sentido em direção à capacidade crescente pode ser retratada como uma seta com capacidade alta para a direita e capacidade baixa para a esquerda."[52] (página 1, tradução nossa).

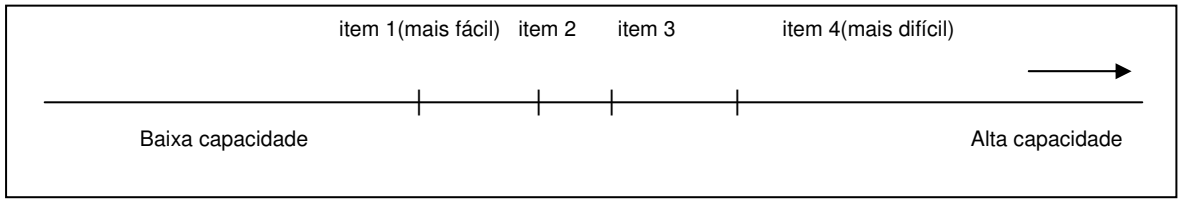


FIGURA 3 – EFINING A VARIABLE
FONTE: WRIGHT E STONE(1979)[53]

Com a linha geral de investigação definida, tem-se então especificado que tipo de capacidade de leitura se pretende medir; seja, a capacidade para encontrar informações específicas, ou a capacidade de encontrar a idéia geral do assunto abordado no texto e/ou quaisquer outras. Essa especificação acontece, normalmente, com a elaboração dos itens do teste, que visam refletir

[52] In order for a test to define a variable of mental ability, the items out of which the test is made must share a line of inquiry. This common line and its direction towards increasing ability can be pictured as an arrow with high ability to the right and low ability to the left."

[53] Ver figura original reproduzida no anexo 5.

o comportamento do candidato para aquela variável. Pode-se dizer que os itens do teste são a definição operacional da variável, ou seja, do construto.

Segundo Alderson et al. (1995, p16-17), "todo teste é uma operacionalização de algumas convicções sobre a linguagem, quer o elaborador se refira a um modelo explicito, quer conte somente com a 'intuição'." (tradução nossa).

Toda teoria é composta principalmente de conceitos psicológicos, isto é de construtos. Assim, os testes podem acessar diferentes construtos juntos ou separadamente, dependendo do propósito do teste. Por exemplo,

> habilidades de síntese e avaliação podem formar parte do nosso construto teórico de leitura. No entanto, como elas são desenvolvidas no nosso teste vai depender muito do propósito para o qual o teste está sendo usado. Tal construto pode ser mais estreitamente definido como a habilidade de identificar, distinguir, comparar e avaliar evidências e opiniões. Mas ele pode bem ser operacionalizado como a habilidade de distinguir entre inferências corretas e incorretas em um item de múltipla escolha baseado em uma passagem curta sobre a história da astronomia, ou pode ser operacionalizado como a habilidade de ler três textos bastante longos, diferentes, apresentando descrições rivais da história da astronomia, os quais o candidato tem que resumir em uma síntese curta.[54] (ALDERSON, 2000, p.118, tradução nossa)

O importante, de acordo com Alderson (2000, p.118) é saber que os construtos "são abstrações que nós definimos para um propósito específico de avaliação" (tradução nossa). Essa definição pode ser, segundo o autor, focada em um ou mais aspectos da habilidade que sejam particularmente relevantes para o propósito do teste.

Bond e Fox sustentam que um dos princípios importantes para o desenvolvimento de medições, por estar implícito nas medidas físicas, talvez

---

[54] "synthesis and evaluation skills may form part of our theoretical construct of reading. However, how these are realized in our tests will depend very much on the purpose for which the test is being used. Such a construct may be more narrowly defined as the ability to identify, distinguish, compare and evaluate evidence and opinions. But it may well be operationalised as the ability to distinguish between correct and incorrect inferences in a multiple-choice item based on one short passage about the history of astronomy, or it may be operationalised as the ability to read three different, fairly long texts presenting competing accounts of the history of astronomy, which the candidate has to summarise in a short synthesis."

não esteja sendo considerado como deveria na definição do construto nas ciências humanas. “Na tentativa de medir os atributos físicos dos objetos, pessoas, ou o tempo, cientistas e também leigos cuidam para medir apenas um atributo do alvo de cada vez.” (2001, p.24, tradução nossa). Esse princípio implica na possibilidade de se medir apenas um atributo separadamente, sem que ele dependa ou sofra a influência de algum outro; isso é denominado *unidimensionalidade*. A importância dessa característica que as medições devem ter é facilmente observável nas medições de atributos físicos de pessoas e objetos, como peso, altura, cumprimento, densidade, entre outros.

Para exemplificar melhor o que significa unidimensionalidade, será reproduzido um exemplo dado por McNamara:

> Vamos considerar o desempenho de crianças em um teste matemático, que consiste de itens de dois tipos básicos: itens em que são usados apenas algarismos (problemas com ‘números isolados’[55]) e itens em que o problema é apresentado em um contexto de mundo real, envolvendo portanto linguagem (problemas ‘com enunciado’[56]). Imagine que estudantes fazendo o teste vêm de duas formações lingüísticas: no primeiro grupo (Grupo A), as crianças são falantes nativas da língua do teste (ex. inglês) e têm poucos problemas em entender a linguagem na qual os problemas com enunciado são apresentados; no segundo grupo (grupo B), as crianças são imigrantes recém chegadas, que estudaram matemática em seu país de origem, mas que têm graus variados de dificuldade com a língua do teste. O desempenho no teste será inconsistente para o último grupo comparado com o primeiro. [...] Claramente, não há uma única coisa sendo medida nesse teste como um todo: ele é um teste de linguagem tanto quanto é um teste de conhecimento matemático.[57] (1996, p.270-1).

---

[55] No original, esses problemas são denominados de *‘naked number’ problems* por McNamara. Foi feita uma tradução livre, por entender não ser necessário levantar, neste momento, a freqüência de utilização desse termo e a respectiva tradução. Também, por considerar a tradução feita suficiente para o entendimento de que esse tipo de problema envolve apenas expressões numéricas descontextualizadas.

[56] Tradução semelhante à anterior, para o termo no original: *‘worded’ problems*. Nesse tipo de problema, entende-se que os dado numéricos são contextualizados por meio de enunciados que propõem uma situação específica.

[57] “Let us consider the performance of children on a mathematics test, which consists of items of two basic types: items where figures only are used (‘naked number’ problems) and items where the problem is presented in a real-world context, thus involving language (‘worded’ problems). Imagine that students taking the test come from two linguistic backgrounds: in the first group (Group A), children are native speakers of the language of the test (e.g. English) and have few

Esse exemplo demonstra, portanto, que o teste não é unidimensional no que diz respeito ao construto, quando se compara os dois grupos. Além disso, o teste não será unidimensional em um sentido de medição, na ausência de algum padrão consistente de dificuldade do item ou habilidade do estudante nos dados obtidos em ambos os tipos de teste. Por exemplo, se dentro de um dos grupos, os alunos obtiverem desempenhos diferentes, compara-se os dois tipos de teste apresentados e considerando a dificuldade dos itens e a habilidade do aluno.

Segundo Pasquali (2003, p.84), os psicólogos afirmam que qualquer desempenho humano é sempre multideterminado ou multimotivado, visto que na execução de qualquer tarefa mais de um traço latente é utilizado. Contudo, admite-se a existência de uma aptidão *dominante*, um fator preponderante que supostamente deve ser medido pelo teste, satisfazendo, assim, o princípio da unidimensionalidade.

Wright estranha o fato de que a imposição de ter uma única variável unidimensional implícita como base os dados cause objeções por parte de alguns estudiosos, uma vez que essa é justamente a essência das medições. Ele assegura também, que "nenhum teste exato pode ser perfeitamente unidimensional. Nenhuma situação empírica pode atingir exatamente os requisitos para medição que geram o modelo Rasch. Este fato da vida é encontrado por cada ciência" (1989, p. 857-860, tradução nossa); visto que até mesmo nas ciências exatas as correções da multidimensionalidade são inevitáveis, não seria razoável esperar exatidão maior nas ciências sociais. O que é possível e deve-se buscar é a tentativa de evitar que a medição contenha um ou mais traços latentes dominantes exercendo influência sobre o

---

problems in understanding the language in which the worded problems are presented; in the second group (group B), the children are recently arrived immigrants who have studied mathematics in their country of origin but who have varying degrees of difficulty with the language of the test. Performance on the test will be inconsistent for the latter group compared with the former. […] Clearly, no single thing is being measured in this test as a whole: it is a test of language as much as it is a test of mathematical knowledge."

traço que se pretende medir (como no exemplo de McNamara acima); e caso isso não seja possível, que a influência seja minimizada ao máximo.

Pasquali (2003, p. 165) afirma que a validade de construto de um teste pode ser trabalhada, entre outros, sob o ângulo da curva de informação da TRI (Teoria de Resposta ao Item), em que foi baseada análise do teste nesta pesquisa e que será apresentada oportunamente. Essa teoria, segundo o autor "trabalha a validade dos testes através de poderosos métodos chamados as funções de informação e de eficiência" em que "a informação fornecida pelo teste é simplesmente a soma das informações fornecidas por cada item do mesmo" e a eficiência "permite comparar a relativa eficiência de um teste com relação a outro em sua capacidade de estimar a aptidão" de quem se submeteu ao teste.

No entanto, antes de entrar na questão da análise do teste segundo a TRI é preciso entender as questões relativas à validade de conteúdo, a seguir e à de critério, mais adiante.

3.2.3 Validade de Conteúdo

Uma definição básica a ser feita em um teste diz respeito ao conteúdo do teste. Quando se pensa em elaborar uma avaliação, a primeira questão é: O que vai ser avaliado? Isto é, qual vai ser o conteúdo da avaliação. Se for uma avaliação de linguagem, ela não pode conter atividade que exijam o conhecimento de cálculos matemáticos, ou fórmulas de física ou química, porque estes conteúdos não fazem parte dos conteúdos estudados na linguagem e, conseqüentemente, não representarão conteúdos válidos para o teste.

De acordo com Brown um teste tem validade de conteúdo se, de fato, representa uma amostra do conteúdo a ser medido e exige do examinando o desempenho de comportamento esperado em relação a este conteúdo. Ele exemplifica o conceito, para melhor entendimento, da seguinte forma:

> Se você está tentando avaliar a habilidade de uma pessoa para falar uma segunda língua em uma situação informal, um teste em que o aprendiz deva responder questões escritas de múltipla escolha que Requerem julgamentos gramaticais não tem validade de conteúdo. Mas um teste que requer que o aprendiz de fato fale dentro de algum tipo de contexto autêntico tem.[58] (2001, p.388, tradução nossa).

Alderson, Clapham e Wall (1995, p. 173) afirmam que "validade de conteúdo envolve 'especialistas' fazendo julgamentos de alguma forma sistemática" e que, em se tratando de "validação de conteúdo nós colhemos julgamentos de pessoas em quem nós estamos preparados para acreditar" (tradução nossa), mesmo quando o nosso julgamento é diferente daquele obtido. Entretanto, eles alertam para o fato de que não é apenas uma questão de escolher as pessoas certas, para que se obtenham julgamentos válidos dos testes, afinal as opiniões dos especialistas podem divergir entre si, o que gera julgamentos inválidos entre si, embora válidos em si mesmos, e acabam por comprometer, conseqüentemente, a validade do teste. Sendo assim, é necessário que se estabeleça um consenso entre eles para que a validade do teste seja, de fato, garantida.

Messick (1994), de acordo com McNamara (1996, p. 16), afirma que a abordagem de avaliação na qual os testes constituem amostras ou simulações da vida real deve considerar as questões de reprodutibilidade e generalização. Isso significa que os resultados dos testes, aplicados em situações similares, devem ser semelhantes e que a partir desses resultados deve ser possível fazer generalizações sobre eles. Ou seja, se o desempenho de um indivíduo em um teste é considerado excelente, em outros testes semelhantes ao primeiro, esse indivíduo deve também obter um desempenho excelente e, além disso, o desempenho obtido no teste deve refletir o desempenho deste indivíduo em situações reais, semelhantes às do teste procedendo a exigência de generalização. Pode ocorrer, inadvertidamente, que um teste que é elaborado com a intenção de medir a capacidade de compreensão de um texto

---

[58] "If you are trying to assess a person's ability to speak a second language in a conversational setting, a test that asks the learner to answer paper-and-pencil multiple choice questions requiring grammatical judgments does not achieve content validity. A test that requires the learner actually to speak within some sort of authentic context does."

acabe medindo na verdade questões lexicais e gramaticais. Nesse caso, esse teste não pode ser considerado válido, pois uma vez que não apresenta os resultados esperados, não é capaz de refletir através dos resultados as informações necessárias para a avaliação do conteúdo pretendido.

No entanto, McNamara afirma que embora a validade de conteúdo seja uma condição necessária para a validade de testes de desempenho, cujas tarefas são amostras ou simulações da realidade, ela não é suficiente. Outros tipos de validade também se fazem necessários: a validade de construto, apresentada no item anterior, e a validade de critério, apresentada a seguir.

3.2.4 Validade de Critério

Para que se possa entender melhor como funciona a validade de critério é preciso que se saiba o significado de *critério* neste contexto de avaliações. Critério, de acordo com o *Dictionary of language testing* (2002, p. 37-8), pode ser:

a) "uma variável externa tal como currículo escolar, julgamento de professor, desempenho no mundo real, ou outro teste." (tradução nossa). O critério é representado no teste devido à impossibilidade de se observar ou medir cada elemento do critério. Sendo assim, o desempenho no teste é usado para predizer o desempenho no critério;
b) "um nível aceitável de conhecimento de, ou desempenho em, um assunto específico de ato de linguagem (por exemplo: inglês para controladores de tráfego aéreo)." (tradução nossa). Neste caso o conceito de critério é mais específico e diz respeito a uma variável do teste;
c) uma qualidade através da qual o desempenho em teste é julgado. "O desempenho bem sucedido em uma atividade de avaliação pode ser caracterizada em termos de critério lingüístico e não lingüístico."

(tradução nossa). No primeiro caso, fluência, coerência e coesão, e no segundo caso adequação do registro de linguagem usado e a completude da tarefa, por exemplo.

De acordo com McNamara, "na testagem, o termo técnico *critério* tem dois significados: (1) desempenho na situação fora do teste; (2) um aspecto no qual o desempenho é julgado, por exemplo, *fluência, adequação"*. (1996, p. 46, nota 4, tradução nossa).

Neste caso, o conceito que interessa é o primeiro proposto por McNamara, cujo objetivo é usar o julgamento que se faz do desempenho de um sujeito em uma situação de teste para predizer qual será o desempenho do sujeito em uma situação real (não de teste). O critério será, então, todos os atos de linguagem relevantes ao desempenho bem sucedido do sujeito na vida real, em relação ao construto definido para e presente no teste.

Mais especificamente em relação à validade de critério, conforme mencionado anteriormente, no final do item 3.2.1, nota 25, Crombach e Meehl (1955) reúnem, sob a classificação de procedimentos de validade de critério-orientado, que será referida aqui apenas como validade de critério, as validades: concorrente e preditiva, exemplificadas da seguinte forma. Quando quer estabelecer a validade de critério,

> O investigador está interessado primeiramente em algum critério que ele deseja prever. Ele administra o teste, obtém uma medida independente do critério, dos mesmos sujeitos, e calcula uma correlação. Se o critério é obtido algum tempo depois que o teste é dado, ele está estudando validade preditiva. Se o escore do teste e o escore do critério são determinados essencialmente ao mesmo tempo, ele está estudando validade concorrente. Validade concorrente é estudada quando um teste é proposto como substituto para um outro (por exemplo, quando uma forma múltipla-escolha de teste de soletração é substituída por tomar ditado), ou um teste mostra ter correlação com algum critério contemporâneo (e.g., diagnóstico psiquiátrico)[59]. (p. 281-382, tradução nossa).

[59] The investigator is primarily interested in some criterion which he wishes to predict. He administers the test, obtains an independent criterion measure on the same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, he is studying predictive validity. If the test score and criterion score are determined at essentially the same time, he is studying concurrent validity. Concurrent validity is studied when one test is proposed as a substitute for another (for example, when a multiple-choice form of spelling test

A fim de melhor entender os exemplos dos referidos autores é interessante observar as definições de validade concorrente e de preditiva dadas pelo Dictionary of language testing (2002, p. 30 e 149). Validade concorrente é “um tipo de validade que diz respeito à relação entre o que está sendo medido por um teste (normalmente um teste desenvolvido recentemente) e outra medida-critério existente” e validade preditiva “mede quão bem um teste prevê o desempenho em um critério externo.” (tradução nossa). Ou seja, faz uma previsão do desempenho provável em um comportamento (usado como critério) no mundo real. A validade preditiva “assume particular importância em testes de proficiência onde o critério pode estar tão distante ou ser tão vago que o próprio teste, por meio da validade de construto, pode ter que justapor tanto o papel de previsor quanto o de critério.” (tradução nossa).

Pasquali (2003, p. 185), afirma que “concebe-se como validade de critério de um teste o grau de eficácia que ele tem em predizer um desempenho específico de um sujeito.” (tradução nossa). Assim como Crombach e Meehl, ele também faz a distinção entre os tipos de validade concorrente e preditiva, como tipos de validade de critério. Para ele, não é tecnicamente relevante se a obtenção da medida foi simultânea (validade concorrente) ou posterior (validade preditiva); a determinação de um critério válido é o que importa. Essa determinação de um critério válido, neste tipo de validade, encontra duas situações diferentes:

a) quando existe um (ou mais) teste comprovadamente validado para a medir algum traço latente – neste caso ele se constitui o critério contra o qual o novo teste pode ser validado com segurança. Segundo este autor esta situação é quase exclusiva de medidas de inteligência.
b) quando não existe nenhum teste comprovadamente validado para medir um traço latente – esta é, segundo o autor, a situação mais comum e, neste caso, a utilização da validação concorrente é

---

is substituted for taking dictation), or a test is shown to correlate with some contemporary criterion (e.g., psychiatric diagnosis).

> precária, porque a validade dos testes existentes é ainda duvidosa. O problema de validar os testes considerando critérios específicos foi motivo de crítica por parte de Cronbach e Meehl (1955). Segundo eles na validade de critério a precisão das inferências só é possível quando os dois testes apresentam uma correlação tão grande que as discordâncias são consideradas insignificantes. Caso contrário, deve-se manter operacionalmente definida cada variável separadamente, ou utilizar a validade de construto.[60]

Pasquali (2003, p. 188) afirma que com esta e outras críticas "a validade de critério deixou de ser a técnica panacéia de validação dos testes psicológicos em favor da validade de construto", porque validar medidas supostamente superiores por medidas inferiores a ela parecia não se justificar. Entretanto, McNamara chama a atenção para um aspecto importante da validade de critério. Ele aponta o fato de que:

> Com relação a testes de desempenho em segunda língua deve-se ter em mente que a linguagem é apenas um dos diversos fatores sendo avaliados. O critério geral é completar de forma bem sucedida uma tarefa na qual o uso da linguagem é essencial. O teste de desempenho é mais que um teste de proficiência básico de competência comunicativa no sentido de que está relacionado a algum tipo de tarefa de desempenho. É perfeitamente possível para alguns examinandos compensar a baixa proficiência de linguagem pela astúcia em outras áreas. Por exemplo, certos traços de personalidade podem ajudar examinandos a obter escores altos em tarefas de interpretação, mesmo que sua proficiência na língua possa estar abaixo do padrão. Por outro lado, examinandos que demonstram proficiência de linguagem geral elevada podem não pontuar tão bem em um desempenho por causa de deficiências em outras áreas.[61] (1996, p. 39, tradução nossa).

---

[60] O texto original diz o seguinte: "But accurate inferences are possible only if the two tests correlate so highly that there is negligible reliable variance in either test, independent of the other. Where the correspondence is less close, one must either retain all the separate variables operationally defined or embark on construct validation."

[61] "With regard to second language performance testing it must be kept in mind that language is only one of several factors being evaluated. The overall criterion is the successful completion of a task in which the use of language is essential. A performance test is more than a basic proficiency test of communicative competence in that it is related to some kind of performance task. It is entirely possible for some examinees to compensate for low language proficiency by astuteness in other areas. For example, certain personality traits can assist examinees in scoring high on interpersonal tasks, even though their proficiency in the language may be

Este fato é particularmente relevante no que diz respeito à escolha do tipo de tarefa escolhida e à correção do teste. Segundo McNamara (1996, p.19), no processo de elaboração de um teste é preciso que se estabeleça a forma como as respostas do teste serão julgadas. Esta forma são os critérios, que freqüentemente referem-se aos construtos psicológicos a serem medidos. O julgamento do desempenho do teste envolve, além da validade, questões relativas à confiabilidade do teste, como se verá a seguir.

3.2.5 Confiabilidade

Em sua definição de confiabilidade do teste Brown afirma que:

> Um teste **confiável** é consistente e fidedigno. As fontes de inconfiabilidade[62] podem estar no próprio teste ou na sua correção, conhecidas respectivamente como confiabilidade do teste e confiabilidade do corretor. Se damos o mesmo teste para o mesmo sujeito ou sujeitos correspondentes em duas ocasiões diferentes, este mesmo teste deveria gerar resultados similares; ele deveria ter a **confiabilidade de teste**.[63] (2001, p. 386, nota e tradução nossa).

Aplicando este conceito no contexto dos testes de suficiência, para que se pudesse afirmar que um teste tem *confiabilidade,* ou que é *confiável,* seria preciso que ele fosse capaz de gerar resultados similares, quando aplicado nos mesmos candidatos, em momentos diferentes (supondo que isso fosse possível e que estes candidatos não tivessem adquirido quaisquer conhecimentos relativos ao domínio avaliado entre as duas aplicações), ou quando aplicado em grupos correspondentes. Um exemplo do segundo caso

---

substandard. On the other hand, examinees who demonstrate high general language proficiency may not score well on a performance because of deficiencies in other areas."

[62] As fontes de inconfiabilidade do teste serão trabalhadas diretamente na análise dos testes piloto e final.

[63] "A **reliable** test is consistent and dependable. Sources of unreliability may lie in the test itself or in the scoring of the test, known respectively as test reliability and rater (or scorer) reliability. If you give the same test to the same subject or matched subjects on two different occasions, the test itself should yield similar results; it should have **test reliability**."

seria se o mesmo teste de suficiência fosse aplicado dois anos seguidos, obtendo resultados semelhantes, uma vez que os grupos de candidatos seriam correspondentes, embora composto por sujeitos diferentes.

Segundo Bachman (1990, p. 24) a "confiabilidade é uma qualidade dos escores do teste" (tradução nossa). Se, por exemplo, o mesmo teste fosse aplicado duas vezes aos mesmos sujeitos e gerasse resultados bem diferentes em cada uma delas, os escores deste teste não poderiam ser considerados confiáveis, porque não produziriam resultados consistentes e não seriam, portanto, indicadores confiáveis da habilidade dos sujeitos avaliados. Da mesma forma, se dois corretores pontuassem com escores bem diferentes os mesmo sujeitos avaliados, o teste não poderia ser considerado confiável, uma vez que, não havendo outras informações, não seria possível decidir qual dos escores adotar. "Confiabilidade conseqüentemente tem a ver com a consistência da medição em diferentes tempos, formas de teste, corretores, e outras características do contexto de medição."[64] (BACHMAN,1990, p. 24, tradução nossa)

Para exemplificar a importância da confiabilidade nos testes de suficiência, supondo que um mesmo teste fosse aplicado em dois momentos e ambientações diferentes para o mesmo grupo de candidatos, não importaria em qual dos dois momentos um indivíduo fizesse o teste, porque o seu escore seria o mesmo. Ou, se fossem elaborados dois ou mais testes de suficiência para serem usados de forma intercalada um a cada ano, não deveria haver diferença para um indivíduo fazer o teste em um ano ou em outro, porque em qualquer um dos testes ele deveria obter o mesmo escore. Bachman e Palmer afirmam que

> confiabilidade é claramente uma qualidade indispensável de escores de teste, pois a menos que eles sejam relativamente consistentes, eles não conseguem nos fornecer absolutamente nenhuma informação sobre a habilidade que nós queremos medir. Ao mesmo tempo, nós precisamos reconhecer que não é possível eliminar completamente as inconsistências. O que nós podemos fazer, no entanto, é tentar minimizar os efeitos das fontes de inconsistências

[64] "Reliability thus has to do with the consistency of measures across different times, test forms, raters, and other characteristics of the measurement context."

> em potencial que estão sob nosso controle, através da elaboração do teste.[65] (1996, p. 24, tradução nossa).

Minimizar as inconsistências dos testes de suficiência é, na opinião desta autora, uma questão não só de necessidade institucional, em função da qualidade dos testes e da garantia de que os candidatos terão o conhecimento de inglês necessário ao desempenho esperado, mas também uma questão de justiça em relação aos candidatos, na medida em que, quanto mais confiável for o teste, mais garantias eles terão de que estarão sendo avaliados justa e adequadamente, independentemente do momento da avaliação.

Com relação à confiabilidade e também à validade de testes é importante ressaltar que nenhuma das duas pode ser obtida de forma absoluta, "no sentido de que nós nunca podemos atingir medidas perfeitamente livres de erros na verdadeira prática, e a adequação de um uso particular da pontuação de um teste vai depender de muitos fatores exteriores a ele mesmo." (BACHMAN, 1991, p.26, tradução nossa). É o avaliador quem vai julgar o quão adequado é um teste e o resultado que ele gera, e quão confiável e válido um teste precisa ser, para o contexto em que ele é usado. Por exemplo, em uma prova de suficiência de inglês em que se queira testar a compreensão de texto, é o elaborador do teste quem vai decidir se uma tradução do texto, ou parte dele, pode ser considerada indício de que o candidato entendeu o texto, ou se isso apenas demonstra a habilidade de decodificar o texto e estabelecer uma correlação entre as duas línguas, sem que o candidato, necessariamente, tenha entendido o significado da mensagem transmitida.

Alderson afirma que não se deve esperar atingir a elaboração de testes perfeitos, ou descobrir qual é o melhor teste de leitura em inglês. A intenção é tentar reunir nesses testes o máximo de qualidades possível, é torná-los, além de convenientes, eficientes e válidos.

---

[65] "Reliability is clearly an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. At the same time, we need to recognize that it is not possible to eliminate inconsistencies entirely. What we can do, however, is try to minimize the effects of those potential sources of inconsistency that are under our control, through test design."

Como visto anteriormente, a correção feita com o estabelecimento de critérios é uma forma de conferir confiabilidade ao teste e, outra forma é assegurar que o teste seja adequado quanto à medida.

3.2.5.1 Adequação quanto à Medida

Para um teste ser confiável ele deve utilizar uma forma de medição adequada aos seus propósitos. Mas, o que se pode ou deve entender por *medida* no contexto de avaliação de leitura em inglês?

De acordo com Pasquali (2003, p.33) a definição de medida afirma que "medir é atribuir números às coisas empíricas" e que esta representação de *coisas* por meio de números depende das características dos objetos focalizados e, portanto, vai apresentar níveis diferentes de qualidade e precisão. Para medir peso ou distância os números são excelente fonte de informações, mas para medir a inteligência eles já não são tão precisos. No entanto, ao contrário do que se poderia pensar, a imprecisão relativa dos números no último caso não significa sua inutilidade nesse tipo de medição.

Bond e Fox (2001, p.1) afirmam que as falhas de qualidade nos experimentos realizados pelas ciências humanas se deve, especialmente, à falta de rigor na medição das variáveis humanas, e que essa falta de qualidade científica afeta diretamente a qualidade das inferências feitas, das decisões tomadas e da política adotada a partir dos resultados dessas medições.

Segundo eles, provavelmente, há duas razões para os pesquisadores das áreas de ciências humanas acreditarem que medições quantitativas como as realizadas nas ciências físicas estão fora do seu alcance: "por um lado, nós superestimamos a qualidade das medições físicas, enquanto por outro lado, nós subestimamos a qualidade das ferramentas que atualmente temos para construir e manter medições objetivas nas ciências humanas". (2001,p. 5, tradução nossa) Esses autores dão alguns exemplos práticos da dificuldade e do tempo necessário para se construir escalas e um deles é o da construção do termômetro, que a seguir apresentamos resumidamente.

Os primeiros registros da tentativa de se construir escalas de temperatura datam do século II DC. Mas só no século XVII cientistas como Galileu obtiveram os primeiros sucessos nessa área, e foi Santorio de Pádua que usou pela primeira vez um tubo de ar invertido em um recipiente com água, cujo nível variava de acordo com a mudança de temperatura. Depois ele calibrou sua escala de temperatura, marcando os níveis da água no fogo e no gelo. As escalas de Celsius e Fahrenheit estabeleceram duas temperaturas conhecidas (congelamento e ebulição) e dividiram a escala em unidades iguais (graus). No século XIX a teoria da termodinâmica de Kelvin passou a ser utilizada como modelo padrão, tendo o hidrogênio como base da medida de calor, por ele ser um gás cujo comportamento é considerado próximo do ideal. E no século XX outros aperfeiçoamentos foram feitos, de tal modo que as medidas de mudanças de temperatura puderam ir além do alcance de precisão da escala Kelvin.

A partir desse tipo de exemplos, Bond e Fox pretendem salientar que mesmo os melhores aparelhos de medição apresentam inconsistências e variações em função do método usado no seu desenvolvimento. O que esses autores procuram enfatizar é que o processo de construção de um sistema de medição que seja reproduzível está sempre sofrendo revisões ao longo do tempo, por não haver um modelo que possa ser considerado *o ideal* e que tenha sido criado isento de imperfeições. De acordo com eles, a despeito das imperfeições, que são inerentes ao processo de criação de escalas de medição, os instrumentos desenvolvidos pelas ciências físicas demonstram a necessidade de se continuar a trabalhar no desenvolvimento dessas escalas.

3.2.5.2 Necessidade de uma Escala de Medição

As avaliações, de maneira geral, devem passar por constantes reformulações, devido a questões relacionadas à validade e confiabilidade dos

testes[66]. No entanto, uma das grandes dificuldades que os profissionais que trabalham na área de avaliação têm que enfrentar é a falta de uma *medida* que permita a padronização dos testes, de forma que eles possam ser elaborados e que seus resultados possam ser julgados com base nos mesmos critérios e que, além disso, os posicione numa mesma escala de classificação.

No caso de testes para o ingresso em programas de pós-graduação em universidades é importante saber se os candidatos, ao longo do tempo, estão sendo testados com base nos mesmos critérios e quais são eles. Supondo que em um teste, exigido em uma universidade como requisito para a obtenção do grau de mestre ou de doutor nos cursos de pós-graduação, a pontuação limítrofe entre 'suficiente' e 'insuficiente' seja sempre 70% de acerto; isto é, o candidato deve ser capaz de realizar corretamente, no mínimo, 70% das atividades propostas para que o seu nível de compreensão de textos seja considerado suficiente para o propósito em questão. Cabe perguntar neste momento: será que a capacidade de um candidato que foi considerada suficiente na prova realizada em um determinado ano é a mesma de um outro candidato, que tenha atingido a mesma pontuação, mas que fez uma outra prova no ano anterior? Se as provas são completamente diferentes e não são equalizadas[67] por meio de uma escala, isto é, não são analisadas por meio de um instrumento que estabeleça qual o grau de dificuldade dos diferentes testes dentro dessa mesma escala, como se pode ter certeza de que os resultados que elas apresentam significam a mesma coisa? O senso comum nos indica que um candidato que demonstre a capacidade de compreensão de 70% de um texto fácil, não terá, necessariamente, a capacidade de compreender 70% de um texto com maior grau de complexidade. Como diz McNamara (1996, p.150) "Um escore mediano em um teste difícil é mais impressionante que um escore alto em um fácil" (tradução nossa).

As questões apresentadas neste exemplo apontam para a relevância de um instrumento de medição que permita, se não resolver totalmente, ao menos minimizar os problemas que as mesmas levantam. É preciso admitir, no

---

[66] Essas questões serão abordadas oportunamente.

[67] O conceito de equalização é visto com mais detalhes na seção 3.5, e pode ser encontrado também no glossário.

entanto, que responder a essas questões não é tarefa simples e elaborar este tipo de instrumento exige tempo.

Através da história do conhecimento humano sabemos, por exemplo, que instrumentos de medição como relógios, termômetros e a escala Richter são exemplos de escalas de medições que levaram muito tempo para serem desenvolvidos, até que fossem bastante confiáveis para serem usados para medir o tempo, a temperatura e a magnitude de abalos sísmicos, respectivamente. Isso deixa evidente que não se deve esperar que escalas de medição nas ciências sociais sejam desenvolvidas em pouco tempo, mas assim como acontece nas ciências físicas, elas também precisam ser constantemente reformuladas e melhoradas até que se tornem suficientemente confiáveis.

Sendo assim, um dos objetivos da tentativa de se criar e utilizar uma mesma escala para as medições nas ciências sociais, neste caso mais especificamente nas avaliações de leitura, é a de possibilitar que se façam abstrações, que possam ir além do dado bruto, assim como acontece nas ciências exatas nos exemplos de medições acima. Uma escala de medição visa, justamente, possibilitar que as análises, generalizações, inferências e tomadas de decisões sejam feitas com mais objetividade e confiabilidade.

De acordo com Bond e Fox, para um modelo de análise de testes ser considerado útil para a investigação de aspectos da condição humana ele precisa apresentar as seguintes propriedades:

> Ele precisa ser sensível à aquisição ordenada das habilidades ou capacidades sob investigação (i.e., ele deveria objetivar a descoberta da ordem do desenvolvimento ou aquisição).
> **Ele deveria ser capaz de estimar as distâncias de desenvolvimento entre as habilidades ordenadas ou pessoas (i.e., ele deveria nos dizer o quanto T é mais desenvolvido, mais capaz, ou mais reabilitado que S)**
> Ele deveria nos permitir determinar se o padrão de desenvolvimento geral apresentado entre itens e pessoas é suficiente para descrever o padrão de desenvolvimento apresentado por cada item e cada pessoa."[68] (2001,p.19, tradução e grifo nossos).

[68] "It should be sensitive to the ordered acquisition of the skills or abilities under investigation (i.e., it should aim at uncovering the order of development or acquisition).

Em outras palavras, é importante que um modelo de análise de testes esteja em alguma escala intervalar, para ser capaz de fazer uma comparação entre os níveis estimados de conhecimento ou desenvolvimento dos indivíduos testados a partir da distância existente entre eles nessa mesma escala. A escala intervalar, dessa forma, vai contribuir para que essas medições científicas nas ciências sociais possam ter o mesmo padrão de qualidade que nas físicas, e que a qualidade das generalizações também possa ser a mesma.

Para Bond e Fox isso significa que as "medições devem satisfazer estes critérios de serem ambas reproduzíveis e aditivas" (2001, p. 2, tradução nossa), por estes serem requisitos básicos para medição de quaisquer atributos científicos quantitativos. Pode-se dar como exemplo prático do que eles querem dizer o fato de que, independentemente do objeto que estiver sendo pesado, 5 kg serão sempre 5 kg; este valor não vai variar na escala de peso de acordo com o objeto medido, ou seja, o valor da medição na escala é reproduzível. Caso se queira adicionar algum valor à essa medida de peso dever-se-á adicionar quilogramas e não litros ou graus Celsius, que não são aditivos entre si, por serem variáveis diferentes e precisarem de escalas de medidas diferentes. Dentro dessa proposta de trabalho, em se tratando de avaliação de leitura entende-se que os resultados obtidos dentro de um mesmo teste ou de um teste em relação a outros são aditivos entre si, por se tratar de uma única variável (compreensão de leitura) [69] e pode-se utilizar a mesma escala de medição na análise dos dados obtidos. Em termos numéricos, os critérios de reprodução e adição também podem ser satisfeitos, pois entende-se que com a utilização de uma escala intervalar, uma nota 07 (sete) vai representar sempre um determinado nível de conhecimento ou capacidade,

---

It should be capable of estimating the developmental distances between the ordered skills or persons (i.e., it should tell us by how much T is more developed, more capable, or more rehabilitated than S).

It should allow us to determine whether the developmental pattern shown among items and persons is sufficient to account for the pattern of development shown by every item and every person."

[69] Não está sendo considerada nesse momento a variação do construto, ou de quaisquer outras que possam influenciar direta ou indiretamente o resultado do teste. Entende-se aqui a variável como uma linha geral de investigação.

independentemente da facilidade/dificuldade dos itens que compõem o teste. Isso significa que os valores dos itens e dos testes podem ser adicionados e o valor da medição passa a ser reproduzível, satisfazendo assim os dois critérios apontados por Bond e Fox como sendo requisitos básicos de medições científicas quantitativas.

3.2.5.3 A Natureza das Medidas e as Escalas

A medida nas ciências psicossociais, segundo Pasquali (2003, p.23) "se insere dentro da teoria da medida em geral que, por sua vez, desenvolve uma discussão epistemológica em torno da utilização do símbolo matemático (o número) no estudo científico dos fenômenos naturais". Apesar de a matemática e a ciência empírica serem sistemas teóricos bem diferentes, terem objetos de estudo e metodologias próprias, é possível e viável uma interação entre esses sistemas de saber distintos. Há muito tempo a ciência percebeu a vantagem de utilizar a linguagem da matemática para descrever seus objetos de estudo. Klein (1974:24, apud Pasquali, 2003, p.25) confirma isso dizendo que "os instrumentos e técnicas de medida propiciam a ponte mais útil entre os mundos do dia-a-dia do leigo e dos especialistas em ciência". Isso é compreensível, se for levada em consideração a dificuldade que um leigo teria de quantificar uma temperatura qualquer e compará-la com outras, sem utilizar uma escala de medidas como referência. Pasquali (2003, p.25) afirma que a teoria da medida, que justifica o uso do número na descrição dos fenômenos naturais "está razoavelmente axiomatizada somente nas ciências físicas, aparecendo ainda lacunar nas ciências psicossociais". No entanto, a idéia de usar números e criar escalas para descrever fenômenos naturais é uma tentativa de tornar estas descrições de coisas abstratas o mais objetivas possível. Pasquali (2003, p.27-29) diz que o principal problema da medida é legitimar a passagem de procedimentos e operações empíricos (observação) para representações numéricas desses procedimentos. Ele explica que a legitimidade do uso de números como descritores de fenômenos naturais só existe se forem

preservadas as propriedades estruturais do número e dos fenômenos naturais neste procedimento. Ou seja, a propriedade de cada número só poderá ser relacionada a um único aspecto dos atributos da realidade empírica. Além disso, a medida deve preservar, se possível, três propriedades básicas do sistema numérico, que são: *identidade,* que diz que um número é idêntico a si mesmo e somente a si mesmo, *ordem,* segundo a qual todo número é diferente do outro em termos de qualidade e magnitude (o que significa que além de diferentes, um é maior que o outro) e *aditividade,* propriedade que garante que os números podem ser somados entre si, de forma que a soma de dois números, excetuado o zero, produz um outro número diferente deles mesmos. Dependendo de quantas propriedades do número são preservadas na medida, resultam diferentes níveis de medida, ou *escalas de medida.*

É importante deixar claro, neste momento, o que entende-se por escala de medida. Sendo assim, as definições são:

Escala (scale): "Um sistema graduado de níveis. A teoria da mensuração faz uso de quatro tipos de escala." (Dictionary of language testing, p.174, tradução nossa).

a) nominal;
b) ordinal;
c) intervalar;
d) proporcional/de proporção.

Entre as classificações de escala, as definições que interessam nesse momento são as de escala nominal, ordinal e intervalar, conforme a descrição que segue:

Escala nominal (nominal scale): "Aquela que consiste na contagem de ocorrências de atributos mutuamente exclusivos. Portanto, é mais a medida da freqüência da ocorrência de um atributo do que o quanto dele está presente." (Dictionary of language testing, p. 128).

> Como seu nome sugere, uma escala nominal compõe-se de números que são usados para 'nomear' as classes ou categorias de um dado atributo. Isto é, nós podemos usar números como código taquigráfico para identificar diferentes categorias. Se nós quantificamos o atributo 'língua nativa', por exemplo, teríamos uma escala nominal. Poderíamos designar diferentes códigos numéricos para indivíduos com língua nativa de origens diferentes, (por exemplo, Amárico = 1, Árabe = 2, Bengali = 3, Chinês = 4, etc.) e então criar uma escala

> nominal para este atributo. Os números designados são arbitrários, uma vez que não faz diferença que número designamos para qual categoria, contanto que cada categoria tenha um único número. A característica distintiva de uma escala nominal é que enquanto as categorias para as quais designamos números são distintas, elas *não são ordenadas* em relação umas às outras. [70] (BACHMAN,1990, p.27, tradução nossa).

No exemplo acima, embora '1' seja diferente de '2', '2' diferente de '3' e assim por diante, os números não são nem mais, nem menos uns que os outros. Esse fato evidencia a propriedade distintiva da escala nominal.

Escala ordinal (ordinal scale):

> Uma escala que ordena objetos de acordo com sua relação uns com os outros. Os pontos na escala ficam em uma relação de 'mais que' ou 'menos que' entre si, [...]. Enquanto uma escala ordinal é capaz de ordenar itens em relação uns aos outros, o tamanho do aumento entre dois pontos adjacentes não pode ser presumido como sendo o mesmo. Uma escala ordinal, conseqüentemente, não pode informar sobre o grau de diferença entre dois itens, por exemplo a diferença de habilidade entre dois candidatos."[71] (DICTIONARY OF LANGUAGE TESTING, p. 137, tradução nossa).

---

[70] As its name suggests, a nominal scale comprises numbers that are used to 'name'the classes or categories of a given attribute. That is, we can use numbers as a shorthand code for identifying different categories. If we quantified the attribute 'native language', for example, we would have a nominal scale. We could assign different code numbers to individuals with different native language backgrounds, (for example, Amharic = 1, Arabic = 2, Bengali = 3, Chinese = 4, etc.) and thus create a nominal scale for this attribute. The numbers we assign are arbitrary, since it makes no difference what number we assign to what category, so long as each category has a unique number. The distinguishing characteristic of a nominal scale is that while the categories to which we assign numbers are distinct, they are *not ordered* with respect to each other.

[71] "A **scale** which orders objects in terms of their relationship to one another. The points on the scale stand in 'more than' or 'less than' relationship to each other, […]. While an ordinal scale is able to order items in relation to each other, the size of the increments between any two adjacent points cannot be assumed to be the same. An ordinal scale, therefore, cannot provide information on the extent of difference between any two **items**, for example the difference in **ability** between candidates."

É possível saber que um candidato cuja classificação é 6 é mais capaz que um outro cuja classificação é 5, mas não é possível dizer que a diferença de habilidade entre estes dois candidatos é a mesma que a existente entre o último deles e um terceiro cuja classificação seja 4.

Sendo assim, esclarece Bachman "o exemplo mais comum de uma escala ordinal é uma classificação, em que indivíduos são categorizados 'primeiro', 'segundo', 'terceiro', e assim por diante, de acordo com algum atributo ou habilidade." (1990, p.28). Portanto, além da propriedade distintiva, a escala ordinal possui a propriedade de ordenação.

Escala intervalar (interval scale): "Uma escala em que os pontos ou unidades de medida são distribuídos em intervalos iguais, como em escalas descrevendo propriedades físicas como altura ou temperatura." (Dictionary of language testing, p.89, tradução nossa).

> Uma escala intervalar é uma numeração de diferentes níveis em que as distâncias, ou intervalos, entre os níveis são iguais. Isto é, além da ordenação que caracteriza escalas ordinais, escalas intervalares consistem de distâncias ou intervalos iguais entre os níveis ordenados. Escalas intervalares, portanto, possuem as propriedades da distinção, ordenação e intervalos iguais. A diferença entre uma escala ordinal e uma escala intervalar está ilustrada na Figura [72] (BACHMAN,1990, p.28)

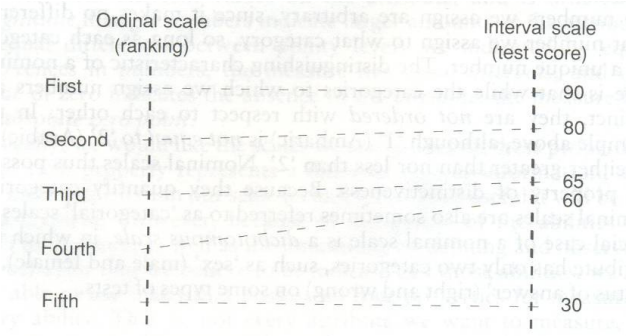


Figure 2.2 Comparison between ordinal and interval scales

FIGURA 3 – COMPARISON BETWEEN ORDINAL AND INTERVAL SCALES.
FONTE: BACHMAN (1990)

[72] "An interval scale is a numbering of different levels in which the distances, or intervals, between the levels are equal. That is, in addition to the ordering that characterizes ordinal scales, interval scales consist of equal distances or intervals between ordered levels. Interval scales thus possess the properties of distinctiveness, ordering, and equal intervals. The difference between an ordinal scale and an interval scale is illustrated in Figure."

Se em um processo de medição, conforme observa Pasquali, apenas a identidade do número é preservada, o processo não chega a ser uma medida, não passa de uma classificação, que resulta em uma escala nominal. Se a medida preserva as três propriedades (identidade, ordem e aditividade) ela resulta em uma escala de razão, que é o tipo de medida mais raro, por ser o mais sofisticado possível. Porém, o que a maioria das medidas procura é preservar pelo menos duas dessas propriedades, que resultarão em uma escala ordinal ou em uma intervalar, dependendo da ordenação dos níveis ocorrer em intervalos diferentes ou iguais. Nas ciências psicossociais se tenta salvar, pelo menos, a propriedade de ordem, que vai nos mostrar que

> se um sujeito resolve corretamente maior número de uma série de problemas do que outro, diz-se que é mais inteligente. Assim, pode-se estabelecer uma escala de inteligência. As inversões que ocorrem são consideradas 'erros de medida' ou de observação, que devem ser tratados dentro da teoria da consistência, a qual visa demonstrar que, apesar desses erros, há consistência na medida. (PASQUALI, 2003, p. 32).

Desse modo, pode-se perceber que consistência na medida é uma característica fundamental da qual dependem as medições de variáveis nas ciências humanas, caso queiram obter maior reconhecimento por sua qualidade científica.

3.3 A QUESTÃO DO ERRO NA MEDIÇÃO: COMO TRABALHAR

Pasquali (2003, p.46) postula que "a medida é um procedimento empírico e não existe procedimento empírico isento de erro". Afirmação "que pode ser empiricamente verificada[73] através de operações de mensuração". Fica mais fácil entender isso, quando se considera que "o número utilizado na medida dos fenômenos naturais não é exatamente o número que os

[73] Fazer esse tipo de verificação não é objetivo deste trabalho, por isso, indicamos ao leitor interessado no assunto a leitura do livro de Pasquali, cuja referência se encontra na bibliografia.

matemáticos estudam" Enquanto *objeto* de estudo da matemática, o número é um ponto, um conceito absolutamente claro e distinto. No entanto, o número utilizado na medição não é um ponto, porque ele não representa especificamente a si mesmo, mas a alguma outra coisa. Esse número é, portanto, um intervalo, "o que significa que ele pode ser mais ou menos ele mesmo, isto é, admite variabilidade, o que é uma maneira elegante de dizer que ele admite erro". (p. 46-7) Esse número, tomado como representação de algo diferente dele mesmo, é o objeto de estudo da estatística, ramo da matemática que serve de base para o modelo teórico a ser utilizado neste trabalho.

A respeito disso, Pasquali propõe a existência de duas fontes possíveis de erro: erros de observação e erros de amostragem. E uma vez que estes erros são inevitáveis, o importante é identificá-los e tentar reduzi-los. Ele divide os *erros de observação* em:

> (1) erros instrumentais devidos a inadequações do instrumento de observação, (2) erros pessoais devidos às diferentes maneiras de cada pessoa reagir, (3) erros sistemáticos devidos a algum fator sistemático não controlado, como por exemplo, medir a temperatura em nível diferente da do mar, e (4) erros aleatórios, que não têm causa conhecida ou cognoscível. (2003, p.47).

Com relação aos *erros de amostragem,* esses ocorrem em virtude da pesquisa empírica geralmente não poder ser realizada com todos os membros de uma população, eventos ou objetos. A seleção de uma amostra fica, automaticamente, sujeita a desvios, isto é, a erros, os quais interferem nas inferências, ou generalização feitas a partir da amostra. Considerando o fato de que o interesse do pesquisador recai sobre a possibilidade de se fazer inferências e generalizações, o erro de amostragem se constitui um problema, uma vez que pode levar a inferências e generalizações errôneas.

Segundo o mesmo autor, o fato de o erro estar sempre presente em qualquer medida faz dele uma constante dificuldade nas tomadas de decisões, baseadas nos resultados de medições. Por isso, é fundamental ter uma idéia o mais aproximada possível da grandeza do erro, para que ele possa ser neutralizado ou ter seus efeitos minimizados.

Diante das dificuldades que as medições apresentam, Pasquali ressalta duas vantagens desse tipo de medição, em relação aos métodos qualitativos ou descritivos. A primeira é a precisão, uma vez que "Apesar da medida nunca ser destituída de erro, ela é capaz de definir limites dentro dos quais os reais valores dos atributos medidos se encontra". A segunda é a simulação, já que

> A manipulação da realidade é geralmente complexa, difícil e custosa. [...] Mas conhecendo com precisão as relações entre os componentes em jogo e suas magnitudes, pode-se utilizar modelos matemáticos para simular os efeitos que queremos estudar e que de outro modo, seria impossível ou impraticável pesquisar. (2003, p. 50-1).

3.4 A TEORIA CLÁSSICA DE TESTES (TCT)

Na medição quantitativa de aspectos do comportamento humano, ou seja, de construtos psicológicos, assim como de qualquer traço da capacidade intelectual do ser humano, a teoria de medida utilizada é chamada Psicometria, teoria de medida que surgiu da necessidade dos cientistas empíricos de utilizar procedimentos mais quantitativos. A Psicometria, segundo Pasquali, "deve ser concebida como um ramo da Psicologia que interfaceia com a Estatística"[74] (2003, p.13-4). Ele diz que "esta teoria trabalha igualmente com dois parâmetros, a saber, a resposta (comportamento) do sujeito e o critério" (p.45-6). A diferença de entendimento do que seja este critério resultou nos dois modelos da Psicometria, que são a Teoria Clássica dos Testes (doravante denominada TCT), na qual o critério é visto como comportamento (futuro) e a Teoria de Resposta ao Item (doravante denominada TRI)[75], que entende o critério como o traço latente.

De acordo com Pasquali

---

[74] Para informações mais detalhadas sobre a Psicometria, ver Pasquali conforme referência bibliográfica.

[75] A TRI será explicada na seção seguinte.

> A Psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens. A Teoria Clássica dos Testes (TCT) se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total (T). Por exemplo, o T em um teste de 30 itens de aptidão seria a soma dos itens corretamente acertados. [...] se [...] o sujeito acertou 20 itens e errou 10, seu escore T seria de 20. (2003, p.67).

Andrade e Valle (1998) dizem que "Resultados obtidos em provas, expressos apenas por seus escores brutos ou padronizados, têm sido freqüentemente utilizados nos processos de avaliação e seleção de indivíduos".

Entende-se que a TCT não é a teoria mais adequada aos propósitos desta pesquisa por dois motivos: o entendimento da autora do que seja critério está de acordo com a TRI e não com a TCT e a intenção não é trabalhar com o escore bruto, porque, entre outros motivos que veremos adiante, ele não reflete a capacidade real do indivíduo.

3.5 O ESCORE BRUTO

Embora seja comum atribuir o tratamento de medição ao escore bruto obtido nos testes, essas pontuações, na verdade, contam mais como observações distintas dos casos, do que como uma medição, em que cada observação representa o resultado verificável da comparação entre o indivíduo e o item. Andrade e Valle (1998, p.13) Os mesmos autores apontam que "os resultados encontrados dependem do particular conjunto de itens que compõem o instrumento de medida, ou seja, as análises e interpretações estão sempre associadas à prova como um todo, o que é a característica principal da Teoria Clássica das Medidas".

A utilização do escore bruto obtido por um indivíduo em um teste de leitura como se forma de medição, implica o entendimento de que os acertos observáveis daquele indivíduo representam a medida da sua capacidade relativamente àquele assunto. Esta é uma inferência bastante simplista, uma vez que os acertos observados num determinado momento não correspondem ao total das capacidades do indivíduo, mas são, na verdade, o ponto de partida

para que se possa estimar a sua capacidade real com relação a determinado assunto. Conforme já foi observado anteriormente, todo processo de medição pressupõe algum tipo de inconsistência, de variação, em maior ou menor grau e que é inerente ao próprio processo de medição, e que por isso mesmo precisa ser considerado.

Pasquali diz que a TCT levanta, então, o seguinte questionamento:

> [...] O que representa este escore do sujeito? Supostamente ele está expressando uma magnitude que seria a magnitude daquilo que o teste queria medir no sujeito. Contudo, toda e qualquer operação empírica, se sabe, é sujeita a erros. Conseqüentemente, este escore bruto do sujeito não pode ser a expressão pura da magnitude daquilo que o teste queria medir no sujeito, mas deve conter igualmente uma porção de erro. (2003, p.68)

Na interpretação de um escore bruto obtido em um teste é preciso primeiro contextualizar este teste para depois poder interpretá-lo. Como exemplifica Pasquali:

> 50 pontos num teste de raciocínio verbal e 40 num de personalidade não oferece nenhuma informação. Mesmo se dissermos que ele acertou 80% das questões não diz muito, visto que o teste pode ser fácil (80% então seria pouco) ou difícil (80% então seria muito). Na verdade, qualquer escore deve ser referido a algum padrão ou norma para adquirir sentido. Uma tal norma permite situar o escore de um sujeito, permitindo: (1) determinar a posição que o sujeito ocupa no traço medido pelo teste que produziu o tal escore e, (2) comparar o escore deste sujeito como escore de qualquer outro sujeito. (2003, p.238-9)

Na TCT o valor atribuído à facilidade de um item é determinado pela proporção de pessoas que acertou aquele item e o valor da habilidade do indivíduo é determinado pela proporção de acertos que ele obteve nos itens do teste. Mas, como se pode observar pelo exemplo de Pasquali, acima, a partir da observação de quantas respostas certas o candidato obteve, ou quantos candidatos acertaram determinado item, não é possível fazer generalizações e estimar a capacidade do candidato, nem a dificuldade do item e muito menos perceber a discriminação que um item consegue fazer entre os candidatos. Isso se deve ao fato de que neste caso não temos um critério de comparação estabelecido para afirmar que o sujeito, em outras circunstâncias acertaria os mesmos itens, itens semelhantes, mais difíceis ou mais fáceis.

A partir disso e visando entender um pouco melhor essa teoria, é preciso considerar um postulado básico da Psicometria Clássica, que diz que "O escore empírico[76] é a soma do escore verdadeiro e do erro". (Pasquali, 2003, p.70) Tem-se assim, nas definições iniciais da teoria, a distinção de três componentes, que são:

> T = escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste
> V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse o erro de medida
> E = o erro cometido nesta medida" (PASQUALI, 2003, p. 69).

Desse modo, Pasquali (2003, p.70) presume que o escore bruto do sujeito difere do verdadeiro devido ao erro. Ele afirma que a tarefa da TCT "consiste em elaborar estratégias (estatísticas) para controlar ou avaliar a magnitude do E", que ocorre devido a fatores estranhos como "defeitos do próprio teste, estereótipos e vieses do sujeito, fatores históricos e ambientais aleatórios". Mas, como não existe uma forma de se determinar o escore verdadeiro do sujeito no teste (V), ou seja, como não se consegue saber a exata pontuação que ele obteria, caso não houvesse erro na medida, a TCT pressupõe que:

1- "*A média do erro é igual a 0*." (porque o erro é considerado um evento aleatório)

2- "*O escore verdadeiro é a expectativa do escore empírico*." (essa expectativa significa que se os erros fossem eliminados, o escore verdadeiro seria igual ao escore empírico)

3- "*A correlação entre o escore verdadeiro e o erro é zero*." (a correlação não existe, porque se considera que os erros sejam aleatórios e não sistemáticos. Se os erros fossem sistemáticos poderiam ser detectados e eliminados)

4- "*Não há correlação entre os erros cometidos num teste qualquer (teste i) e num teste paralelo (teste j)*." (testes paralelos são aqueles que

[76] Pasquali usa os termos 'escore empírico' e 'escore bruto' como sinônimos.

medem a mesma coisa, por meio de itens (tarefas) diferentes – os erros cometidos em diferentes testes são, portanto, independentes uns dos outros)

5- "*Os escores verdadeiros de testes paralelos são iguais.*" Ou seja, "*A variância de um teste é igual à variância de um teste paralelo.*"[77] (Mas, "quando os escores verdadeiros são iguais em ambos os teste, mas as variâncias dos erros são diferentes, [...], os testes estão medindo a mesma coisa, mas produzindo diferente variabilidade")

Considerando o que foi exposto, pode-se entender porque os escores brutos dão pouca ou nenhuma informação sobre o conhecimento do indivíduo testado; são escores quase sem significado por si só e que, por isso, são normalmente convertidos em outros tipos de escores para fins de interpretação.

3.6 O ESCORE PADRONIZADO E PERCENTIL

Atualmente, há outras duas formas de se calcular e expressar os resultados obtidos em testes, as quais, além de serem muito utilizadas, possibilitam julgamentos mais significativos que o escore bruto: o escore padronizado (também chamado escore *z,* ou escore padrão) e o escore percentílico (ou percentil).

O escore padronizado interpreta um escore bruto, marcando a que distância ele se encontra da média do grupo do qual aquele escore faz parte. Essa distância é indicada em unidades de desvio padrão para aquele grupo de referência. Pode-se saber, dessa maneira, quantos desvios padrão acima ou abaixo da média o escore de um indivíduo está. Os pesquisadores dizem que, em geral, aproximadamente dois terços de todos os escores de um teste vão estar entre -1 e +1 desvio padrão da média. O outro um terço vai estar distribuído igualmente acima e abaixo desses dois terços.

Martin, no artigo *Standardized Scores* afirma que para entender o que é o escore padronizado

---

[77] Dizemos que a variância de um teste é igual, quando a distribuição dos erros em ambos é igual. Isso não significa que os erros sejam os mesmos.

> a chave é a palavra “classificações”, porque escores padronizados são especificamente planejados mais para mostrar uma classificação de resultados do que um nível de resultados. O escore padronizado mostra como um escore particular se compara a todos os outros escores daquele teste, mas ele não mostra nenhuma informação sobre o nível de resultado do escore. (2008, p.1) [78]

Segundo Siegle, no artigo intitulado *Interpret Raw Scores - Standardized Scores* (p.1) “Para calcular um escore *z*, subtrai-se a média do escore bruto e divide-se o resultado pelo desvio padrão. (isto é, escore bruto = 15, média = 10, desvio padrão = 4. Portanto 15 menos 10 é igual a 5. 5 dividido por 4 é igual a 1.25. Então, o escore *z* é 1.25.)” Há, porém, duas dificuldades que um escore *z* apresenta, “a saber: (1) a presença de escores negativos, pois o *z* vai de menos infinito a mais infinito (mais praticamente, de -5 a +5) e (2) a presença de decimais”. (Pasquali, 2003, p.247) A maneira de se evitar estas duas dificuldades é multiplicar o *z* por um coeficiente e somar este produto a uma constante, sendo ambos os valores arbitrários (coeficiente e constante).

Um grande problema que Pasquali aponta em relação ao escore padronizado é ligado à

> amostra utilizada para fornecer dados empíricos sobre os quais serão efetuados os cálculos para a produção das normas. Esta amostra, dita grupo normativo, deve ser estatisticamente representativa da população; o que implica, praticamente, em que os grupos normativos são geralmente constituídos de números grandes de sujeitos, implicando em custos proibitivos que tipicamente assustam os pesquisadores ao quererem enveredar em projetos de padronização de testes. (2003, p.249).

O escore percentílico indica a porcentagem de sujeitos em um determinado grupo que estão abaixo do escore de um candidato. De acordo com Pasquali (2003, p.242) a grande dificuldade da escala percentílica é que ela é uma escala ordinal, ou seja, as distâncias entre escores sucessivos não

---

[78] “The key is the word “ranks” because standardized scores are specifically intended to display a ranking of achievement rather than a level of achievement. The standardized score displays how an individual score compares to all the other scores from that test, but it does not display any information about the achievement level of the score.”

são constantes, eles variam de acordo com a posição em que o escore estiver, no início/fim da escala ou no meio dela, porque os intervalos entre os percentis dos extremos da escala são maiores que os dos medianos. Outra dificuldade é que o percentil de um candidato é relativo a um determinado grupo apenas. Esse mesmo sujeito, se inserido em um grupo diferente terá um percentil diferente do anterior, porque o nível dos candidatos nesse segundo grupo vai variar em relação ao primeiro. Portanto, fica também difícil comparar dois grupos diferentes utilizando esse tipo de escore, uma vez que a escala ordinal resultante de um teste em um determinado grupo será diferente da resultante em outros grupos e apresentará variação de distância entre escores.

Apesar de a TCT produzir testes de qualidade, ela apresenta algumas limitações. Segundo McNamara (1996, p.152) os valores dos dados analisados com a Teoria Clássica são instáveis, porque ela não tem como compensar a inevitável variabilidade existente entre os diferentes grupos testados. Da mesma forma, a estimativa da capacidade de um candidato é limitada pelas características dos itens escolhidos para o teste, como sua dificuldade.

Pasquali apresenta quatro limitações, que foram salientadas por Hambleton, Swaminathan e Rogers (1991) e que serão reproduzidas resumidamente a seguir.

> 1) Os parâmetros clássicos dos itens (dificuldade e discriminação) dependem diretamente da amostra de sujeitos utilizada para estabelecê-los (*group-dependent* ou *sample-dependent*). Daí, se a amostra não for rigorosamente representativa da população, aqueles parâmetros dos itens não podem ser considerados válidos para esta população.
> 2) A avaliação das aptidões dos testandos também depende do teste utilizado (*test-dependent*). Assim, testes diferentes que medem a mesma aptidão irão produzir escores diferentes da mesma aptidão para sujeitos idênticos. Testes com índices de dificuldade diferentes evidentemente produzirão escores diferentes.
> 3) A definição do conceito de fidedignidade ou precisão na teoria clássica dos testes [...] é concebida como a correlação entre escores obtidos de formas paralelas de um teste ou, mais genericamente, como o oposto do erro de medida. [...] é praticamente impossível satisfazer as condições de definição de formas paralelas e, no caso do erro de medida, é postulado que este seja idêntico em todos os examinandos, postulado improvável (Lord, 1984), uma vez que fica difícil presumir que sujeitos de baixa aptidão, por exemplo, cometam erros iguais aos de habilidade superiores.
> 4) ... teoria clássica dos testes [...] é orientada para o teste total e não para o item individual. Toda a informação do item deriva de considerações do teste geral, não se podendo assim determinar como o examinando se comportaria diante de cada item individual.

> Ademais, a análise de cada item é feita em função do escore total, do qual cada item faz parte. Então, fica um tanto incongruente avaliar a qualidade do item quando ele próprio contribui para a mesma. (2003, p. 80-1).

As limitações apresentadas acima entre outras, levaram à busca de outras teorias que permitissem estabelecer:

> 1) características do item sem ser dependentes da amostra de sujeitos utilizados;
> 2) escores dos examinados independentes do teste utilizado;
> 3) um modelo ao nível do item em vez do teste, de sorte que a análise do item não dependa dos demais itens do teste;
> 4) um modelo que não exija formas rigorosamente paralelas para avaliar a fidedignidade;
> 5) um modelo que ofereça uma medida de precisão para cada nível de aptidão, isto é, que a avaliação da aptidão tenha igual exatidão em todos os seus níveis e não somente nos níveis medianos como faz a psicometria clássica. (PASQUALI, 2003, p.82)

3.7 A TEORIA DE RESPOSTA AO ITEM (TRI) E O MODELO RASCH

Bond e Fox (2001) afirmam que já no começo do século 20 havia a tentativa de se construir escalas de medições com intervalos iguais, nas ciências sociais, mas que essas escalas não eram objetivas[79]. Dessa tentativa surge a teoria moderna da Psicometria, a Teoria de Resposta ao Item (TRI).

Conforme mencionado anteriormente, a TRI difere da TCT por trabalhar com traços[80] latentes (aptidões, habilidades, etc). Isso implica no fato de que enquanto a TCT visa produzir *testes* de qualidade, a TRI visa produzir *itens* de qualidade. De acordo com Pasquali (2003, p. 67) a TRI não se interessa pelo escore total do teste, mas por cada item de que ele se compõe.

---

[79] Um dos modelos da TRI é o Modelo Rasch, que será utilizado no desenvolvimento deste trabalho. Segundo esses autores, nesse momento o modelo Rasch é a única ferramenta que proporciona a desejada objetividade na construção de escalas de medição nas ciências sociais.

[80] De acordo com o "Dictionary of Language Testing" *traços* são aspectos ou características de uma pessoa, que subjazem e explicam o seu comportamento.

Além disso, essa teoria procura saber a probabilidade acerto e erro de cada item e quais os fatores que afetam essa probabilidade.

Na prática, o que a TRI faz é apresentar ao sujeito uma série de estímulos, tais como itens de um teste, aos quais ele vai responder. As respostas dadas são, então, analisadas e, a partir da análise pode-se fazer inferências sobre o traço latente do sujeito e levantar hipóteses sobre a relação entre as respostas observadas e o nível do traço latente desse sujeito.

A TRI é uma teoria que serve de base para que se possa colocar em uma mesma escala de classificação as pessoas avaliadas e os itens de um teste, a partir das respostas dadas a esses itens. Além disso, ela facilita o trabalho envolvido na equivalência, elaboração e análise de testes, bem como a manutenção de todos eles em uma mesma escala de classificação.

Há duas formas de equivalência[81] que nos interessam em especial. A primeira é a que chamamos simplesmente *equivalência* ou *equalização*, que se refere à equivalência interna de um teste. Tal processo de comparação é feito entre itens individuais, entre tarefas ou testes inteiros, com o intuito de estabelecer a equivalência interna de algum(ns) item(ns), de alguma(s) tarefa(s) ou do teste todo. A segunda é a equivalência entre diferentes testes, chamada de *equivalência de teste,* quando a equivalência é feita entre duas ou mais formas de um mesmo teste, isto é, testes construídos a partir da mesma especificação, a fim de que as mesmas habilidades sejam medidas; ou *equalização de teste,* quando no processo são comparadas as dificuldades de dois ou mais testes, para que os escores desses testes possam ser equivalentes e fazer parte de uma mesma escala, permitindo assim a comparação entre os testes.

Simplificando o que encontramos em Pasquali (2003, p.261) sobre esse conceito, vemos que a equalização serve para converter um sistema de unidades de medidas em outro, como acontece na conversão das unidades de medida de temperatura, quando passamos de Celsius para Fahreinheit e vice-versa. A equalização busca ajustar os diferentes níveis de dificuldade entre formas diferentes de um mesmo teste ou entre testes diferentes, que

[81] O termo *equivalência* é considerado sinônimo de *equalização*, segundo o "Dictionary of Language Testing".

teoricamente foram criados para ter um nível de dificuldade igual, mas que na prática não o têm.

De acordo com o *Dictionary of language testing*,

> um método de se estabelecer a equivalência de duas formas de testes é aplicar ambos ao mesmo grupo de sujeitos de teste ou avaliados; alternativamente, se a análise TRI é usada, diferentes formas de combinações contendo alguns itens ou atividades comuns pode ser aplicado a diferentes grupos de examinandos.[82] (2002, p. 199).

Pasquali (2003, p.84) afirma que duas das suposições da TRI são especialmente relevantes, a unidimensionalidade (assunto já abordado no item 2 acima) e a independência local. Este último postulado, segundo o autor, afirma que

> mantidas constantes as aptidões que afetam o teste, as respostas dos sujeitos a quaisquer dos itens são estatisticamente independentes. Isso quer dizer que os itens são respondidos em função do traço latente predominante e não em função de memória ou outros traços latentes. [...] Assim, a independência local significa que, para examinandos com uma aptidão dada, a probabilidade de resposta a um conjunto de itens é igual aos produtos das probabilidades das respostas (produtório) do examinando a cada item individual. (PASQUALI, 2003, p. 85).

Podemos dizer, em outras palavras, que se forem controlados os outros fatores que afetam o teste (traços latentes secundários), a única fonte de variação será o próprio traço latente. Assim as respostas tornam-se independentes e com a independência local das respostas temos também a unidimensionalidade do teste, uma vez que os itens analisados estarão medindo o mesmo traço latente, em função do qual são dadas as respostas de cada item.

---

[82] Há outros métodos de se estabelecer a equivalência de escores e de testes, que serão abordadas na dissertação.

"One method of establishing the equivalence of two test forms is to administer both to the same group of Trial subjects or test takers; alternatively, it IRT analysis is used, different composite forms containing some common items or tasks may be administered to different groups of test takers."

Pasquali (2003, p.91) diz que na TRI, “o desempenho do sujeito numa tarefa [...] depende de: (1) aptidão do sujeito [...] e (2) dos parâmetros dos itens” [...] Daí, a primeira tarefa da TRI é viabilizar modelos que possam permitir a descoberta dos parâmetros dos itens.” A estimação do parâmetro é feita com base em dados empíricos. Por exemplo, em um jogo de par ou ímpar, de 100 jogadas feitas, 70 dão par e 30 ímpar. Podemos dizer que a probabilidade mais verossímil de que uma jogada dê par é de 70/100, ou seja, é de 0,70. Esse é um método de avaliação chamado de máxima verossimilhança, porque segundo os dados empíricos obtidos, os valores estimados são os mais plausíveis.

A TRI possui modelos logísticos que se distinguem pelo número de parâmetros que utilizam para descrever um item. Há modelos de um, dois e três parâmetros. O modelo Rasch é um modelo logístico de um parâmetro, criado pelo matemático e estatístico Georg Rasch e publicado pela primeira vez em 1960. “Ele permite que sejam feitas estimativas da capacidade implícita do candidato, analisando o desempenho do candidato em um grupo de itens, após terem sido feitas considerações sobre a dificuldade dos itens e do quão bem eles se ajustam ao nível de capacidade do candidato.” (MCNAMARA, 1996, p.152, tradução nossa).

Na prática, a partir do momento em que se estabelecem valores para quaisquer eventos ou para as observações desses eventos, uma abordagem quantitativa está sendo usada, mesmo que ela seja marcada apenas com a presença ou ausência do evento ou observação. Esses valores estabelecidos são os escores brutos. Em lugar de usá-los como estimativa da capacidade de uma pessoa no teste, o modelo Rasch transforma o escore bruto de uma escala ordinal em uma escala intervalar. Este modelo proporciona, portanto, a base necessária para a mensuração de traços e atributos[83] quantitativos, com

[83] Segundo o Dicionário Houaiss *atributo* significa “o que é próprio e peculiar a alguém ou a alguma coisa”; na estatística ele é um “aspecto, qualitativo ou quantitativo, que distingue um integrante de um conjunto observado”. Um *atributo* é um valor específico de uma variável. Por exemplo, a variável *sexo* ou *gênero* tem dois atributos: masculino e feminino. E uma *variável* é “um atributo ou traço de pessoas ou coisas que pode assumir diferentes valores.” (Dictionary of Language Testing)

base em escores brutos obtidos da interação entre pessoas e itens, transformando uma média de acerto em probabilidade de acerto individual.

Tanto na análise tradicional, quanto na análise do modelo Rasch são determinadas as características de um determinado grupo de pessoas em relação a um determinado conjunto de itens e as características dos itens em relação àquele grupo de pessoas. A diferença é que, ao contrário do que acontece no modelo Rasch, na análise tradicional não temos como saber se essas características de capacidade das pessoas e dificuldade dos itens serão mantidas para as mesmas pessoas em relação a itens diferentes dos primeiros, e para os itens, se eles forem testados em outro grupo de pessoas. (MCNAMARA, 1996, p.153).

Wright exemplifica esse problema de forma bastante clara:

> Se todo um grupo específico de itens foi testado por uma criança que você quer medir, então você pode obter a posição do seu percentil entre quaisquer grupos de crianças em que foi usado o teste padronizado. Mas como você interpreta esta medida além dos limites daquele grupo de itens e daquelas crianças? Mude as crianças e você tem uma nova medida. Mude os itens e você tem uma nova medida outra vez. Cada conjunto de itens mede uma capacidade em si mesmo. Cada medida depende para o seu significado da sua família de avaliados. Como podemos fazer medições mentais objetivas e construir uma ciência de desenvolvimento mental quando trabalhamos com medida de elástico?"[84] (1967, p.1, tradução nossa).

Segundo ele, para conseguir mais objetividade nas medições é preciso satisfazer duas condições: primeiro, que a calibragem dos instrumentos de medição seja independente dos objetos que forem usados na calibragem,[85]

---

[84] "If all of a specified set of items have been tried by a child you wish to measure, then you can obtain his percentile position among whatever groups of children were used to standardize the test. But how do you interpret this measure beyond the confines of that set of items and those groups of children? Change the children and you have a new yardstick. Change the items and you have a new yardstick again. Each collection of items measures an ability of its own. Each measure depends for its meaning on its own family of test takers. How can we make objective mental measurements and build a science of mental development when we work with rubber yardstick?"

[85] **Calibragem (calibration):** "A calibragem de um teste envolve a determinação do valor dos itens do teste em uma escala particular de medição, em outras palavras ela reflete a dificuldade do item. [...] O processo de calibragem de itens dá significado aos valores da escala em termos

segundo, que a medição dos objetos seja independente do instrumento usado para medir. Embora na prática estas condições só possam ser *aproximadas*, é essa aproximação que faz com que a medição seja objetiva.

Wright e Linacre consideram a visão que Rasch teve desse problema simples, mas profunda.

> Primeiro, ele percebeu que, para ser de alguma utilidade, uma medida deve conservar seu *status* quantitativo, dentro do possível, independentemente do contexto em que ela ocorra. Para uma medida ser útil para medição, ela deve manter seu comprimento de calibragem independente do que ela está medindo. Assim também, cada teste ou item da escala de classificação deve manter seu nível de dificuldade, não importando quem está respondendo a ele. Segue-se que a pessoa medida deve conservar o mesmo nível de competência ou capacidade independente de quais itens de teste em particular são encontrados, contanto que quaisquer itens que sejam usados pertençam ao grupo de itens calibrados que define a variável em estudo. A implementação desse essencial conceito de invariância ou objetividade tem se estendido com sucesso à condescendência (ou severidade) de corretores e à estrutura de medida de escalas de classificação.[86] (1989, p. 859).

De acordo com vários autores, Rasch estabeleceu este princípio chamado *invariância* em seu modelo com base no fato de que as comparações invariantes são características das medições na física. Além disso, sua

---

de conhecimentos ou habilidades requeridas para obter sucesso em atividades dentro de uma gama particular de valores." (Dictionary of language testing – p.18)

[86] "First, He realized that, to be of any use at all, a measure must retain its quantitative status, within reason, regardless of the context in which it occurs. For a yardstick to be useful for measurement, it must maintain its length calibrations irrespective of what it is measuring. So too, each test or rating scale item must maintain its level of difficulty, regardless of who is responding to it. It also follows that the person measured must retain the same level of competence or ability regardless of which particular test items are encountered, so long as whatever items are used belong to the calibrated set of items which define the variable under study. The implementation of this essential concept of invariance or objectivity has been successfully extended in the past decade to the leniency (or severity) of raters and to the step structure of rating scales." (artigo sem paginação).

estrutura formal possibilita a separação algébrica dos parâmetros dos indivíduos e dos itens no processo de cálculo das estimativas estatísticas. O fato de que o modelo Rasch trabalha sob uma perspectiva probabilística e não determinística proporciona às mensurações nas ciências sociais uma confiabilidade mais análoga à das ciências exatas.

Outra vantagem de utilizar o escore probabilístico em lugar do escore bruto/observado é que a calibragem do teste, ou do item, pode ser feita através da estimativa da dificuldade do(s) item(ns), com a sua localização em um continuum. Isso significa, na prática, a possibilidade de construir uma escala estimativa de avaliação que possa ser equalizada.

3.8 DIFICULDADE DO ITEM

Uma escala de medidas para avaliações de habilidade de leitura tem como principais objetivos permitir a estimativa da real capacidade do candidato e comparar a capacidade dele em relação aos demais.

A importância de se estabelecer a dificuldade de um item é que essa dificuldade tem como conseqüência a discriminação que o item é capaz de fazer entre os candidatos em uma seleção ou avaliação, alocando-os em diferentes níveis de habilidade.

Como mencionamos anteriormente, a real capacidade do candidato é estimada *a partir* dos seus acertos, mas ela não é o escore bruto obtido na avaliação.

Da mesma forma, não podemos fazer a análise de um item com base apenas no número de candidatos que acertaram e que erraram cada questão. A análise de um item deve poder determinar a facilidade/dificuldade do item, que é dada pela proporção de pessoas que acertaram/erraram o item. No entanto, neste caso, não é apenas o número de pessoas que entra em jogo, mas também a capacidade de cada pessoa, uma vez que um item considerado difícil para alguns, pode ser fácil para outros e vice-versa. Além disso, o procedimento de análise do item deve ser capaz de mostrar qual a contribuição desse item na discriminação entre os candidatos, isto é, a análise do item deve

informar se a capacidade de cada candidato é consistente quando comparada com a capacidade que ele demonstra em outros itens do teste.

Segundo Bond e Fox, a TCT sustenta que a dificuldade de um item é definida pela proporção de pessoas que acertam o item. Eles explicam de forma bastante clara quais as implicações disso, da seguinte maneira:

> Esta definição isolada nos diz que a dificuldade de um item depende diretamente da distribuição das habilidades das pessoas que responderam ao item. Imagine esta lógica nas ciências físicas, dizer a uma pessoa que a altura de 6 pés em uma régua depende do quê a pessoa está medindo!
> A teoria tradicional de teste nas ciências sociais, portanto, confunde a calibragem do item e a medição do atributo. Instrumentos de medição devem primeiro ser criados e as unidades calibradas, para que nós todos concordemos com a reprodutibilidade das suas localizações. Só então estaremos justificados ao usar estes instrumentos para medir o quão alta, pesada, ansiosa, inteligente, ou estressada uma pessoa é."[87] (2001, p. 2-3, tradução nossa).

Pasquali afirma que há um problema teórico no cálculo do índice de discriminação com base no escore total do teste, porque para analisar a discriminação entre os candidatos que um item é capaz de fazer em um teste, o que se toma por base são as informações obtidas a partir dos outros itens do teste, ou seja, o escore total. No entanto, na medida em que ainda é preciso determinar a adequação de todos os outros itens, não é possível saber se o teste é de fato unidimensional, condição necessária para que se possa obter e usar o escore total da maneira mencionada. Uma forma de resolver esse problema é verificar se os itens referem-se ao mesmo fator. Porém, a informação que a TCT consegue obter sobre o índice de discriminação de itens muito fáceis ou muito difíceis não é confiável, por se aproximar de zero.

---

[87] "This definition alone tells us that the difficulty of an item depends directly on the distribution of the abilities of the persons who responded to the item. Imagine this logic in the physical sciences, telling a person that the height of 6 feet on a ruler depends on what the person is measuring!
Traditional test theory in the social sciences thus confounds the item calibration and the measurement of the attribute. Measurement instruments must first be created, and the units calibrated, so that we all agree on the reproducibility of their locations. Only then are we justified in using these instruments to measure how tall, heavy, anxious, bright, or stressed a person is."

McNamara exemplifica a questão da facilidade/dificuldade do item da seguinte forma:

> Se os itens são muito fáceis, então pessoas com níveis diferenciados de capacidade ou conhecimento vão acertá-los, e as diferenças de capacidade ou conhecimento não serão reveladas pelo item. Da mesma forma, se os itens são muito difíceis, então candidatos mais ou menos capazes vão errá-los igualmente, e o item não vai nos ajudar na distinção entre eles."[88] (2000, p.60, tradução nossa).

A TRI diz que a dificuldade do item é diretamente proporcional ao nível ou tamanho do traço latente necessário para que este item seja acertado. Quanto maior o nível do traço necessário para o acerto do item, mais difícil é o item.

De acordo com Pasquali (2003, p.139) na TRI "Discriminação se refere ao poder de um item em diferenciar sujeitos com magnitudes diferentes de traço do qual o item constitui a representação comportamental. Quanto mais próximas forem as magnitudes do traço que um item puder diferenciar, mais discriminativo ele é". Como na TRI a dificuldade do item não interfere no cálculo da sua capacidade de discriminação, a estimação desse parâmetro para itens muito fáceis ou difíceis demonstra ser mais adequada nessa teoria.

3.9 CORREÇÃO DE TESTES

Em avaliações de linguagem, quaisquer que sejam, existem diversas fontes de variabilidade que podem afetar direta ou indiretamente os resultados e sua análise. Em se tratando de avaliações de desempenho algumas fontes de variação se refletem diretamente nos resultados obtidos. McNamara aponta três fontes de variabilidade em uma avaliação de escrita como exemplo, mas

---

[88] "If the items are too easy, then people with differing levels of ability or knowledge will all get them right, and the differences in ability or knowledge will not be revealed by the item. Similarly, if the items are too hard, then able and less able candidates alike will get them wrong, and the item won't help us in distinguishing between them."

que também se aplica ao tipo de avaliação de leitura em inglês que está sendo proposto neste trabalho. São elas:

> Primeiro, a habilidade relativa dos candidatos vão diferir, e a não ser que o teste suponha uma tarefa simples dentro da capacidade de todos os candidatos (ou uma difícil além da sua capacidade), esta variação na habilidade será refletida nos escores. Segundo, poderá haver variabilidade associada com a tarefa: se há uma escolha de tarefas, então candidatos poderão obter diferentes escores dependendo de qual tarefa eles tiverem escolhido. Terceiro, há variabilidade associada com os corretores, tanto que se um candidato tiver um corretor diferente, ele ou ela podem ter obtido um escore diferente para o mesmo desempenho."[89] (1996, p. 121-2, tradução nossa).

3.9.1 Variabilidade relacionada ao leitor

Segundo Alderson (2000, p. 33), várias pesquisas têm investigado possíveis fontes de variabilidade entre leitores. Entre elas, ele cita o conhecimento, a motivação do leitor, os propósitos de leitura, as estratégias de leitura e as características pessoais de cada leitor, tanto físicas quanto psicológicas, que podem influenciar de alguma forma o desempenho deste leitor. Algumas dessas fontes influenciam mais diretamente que outras e são mais relevantes para esta pesquisa, em função da influência que exercem sobre certos aspectos dos resultados obtidos em testes de leitura compreensiva.

Existem muitos detalhes possíveis de serem observados e analisados a este respeito, alguns dos quais foram abordados, juntamente com as questões de leitura e do leitor no segundo capítulo e, embora o objetivo, no

---

[89] "First, the relative ability of the candidates will differ, and unless the test involves a simple task within the competence of all candidates (or a difficult one beyond their competence), this variation in ability will be reflected in the scores. Secondly, there may be variability associated with the task: if there is a choice of task, then candidates may gain different scores depending on which task they have chosen. Thirdly, there is variability associated with raters, so that if a candidate had had a different rater, he or she might have gained a different score for the same performance."

momento, não sejam esses detalhes especificamente, é importante mostrar um exemplo prático relacionado à variação do conhecimento, pois ela é, necessariamente, considerada nas correções de testes de leitura.

Não raro pode ocorrer de o candidato compreender o texto, mas não ser capaz de se expressar, conforme o esperado, por meio de um texto escrito, isto é, de redigir uma resposta discursiva adequada, segundo os padrões estabelecidos pelo elaborador e pelo corretor; já outro candidato, que não tenha compreendido tão bem o texto quanto o primeiro, pode ter habilidade para redigir uma resposta, que seja considerada mais adequada, usando seu conhecimento de mundo e sua habilidade discursiva.

Para perceber melhor a redação da resposta como uma fonte de variação é preciso ter em mente, em primeiro lugar, que por menor que seja a resposta, ela deve ser coerente, tanto em relação ao texto quanto ao contexto em que ela é produzida. Isto significa que, no teste de leitura, a resposta deve conter as informações requisitadas, de acordo com o texto lido e apresentar essas informações de uma forma que possa ser compreendida pelo leitor alvo, neste caso, pelo corretor. Em segundo lugar é preciso considerar que quanto mais longa for a resposta maior deve ser a sua *textualidade*, que segundo Costa Val (1999, p.5) é como se chama um "conjunto de características que fazem com que um texto seja um texto, e não apenas uma seqüência de frases", ou seja, a resposta, em si, também constitui um texto. Podemos ver como a falta de textualidade interfere na correção, conforme o exemplo abaixo.

Situação:

A questão nº2 reproduzida abaixo foi proposta em um teste para alunos do curso de Compreensão de Textos em Inglês do CELIN[90]. A grande maioria desses alunos freqüenta este curso com o objetivo de melhorar a sua leitura para passar no teste de suficiência para mestrado ou doutorado. As respostas foram transcritas fielmente das provas dos alunos em questão, sem alterar acentos, pontuações, ou mesmo a diagramação das respostas nas folhas de prova. A parte do texto em que se encontra a resposta esperada é a seguinte:

[90] Centro de Línguas e Interculturalidade da UFPR.

Texto:

All that said, the odds of getting a really good raise, say 20%, are very small if you stay in your current job, no matter what. The sad truth is that you almost always have to change bosses, departments, or companies to get a really good raise. They figure that you were working for a certain salary, so they shouldn't suddenly have to give you a lot more to do the same thing. The answer to that, of course, is that they need you and you have a better offer elsewhere. That's why they should suddenly give you a lot more.

Questão 2:

Segundo o autor, por que as chances de conseguir um aumento no mesmo emprego são pequenas? Explique.

Comentário sobre a questão: A questão poderia parar no ponto de interrogação. No entanto, para evitar uma tradução literal, que não era desejada neste teste, a explicação foi solicitada, para que o aluno tentasse usar suas próprias palavras.

Resp. 1:

É que eles precisam de você (no caso a empresa precisa) e você uma mudança, um aumento de salário, ou promoção. É porque eles poderiam de repente te dar um aumento, mas as chances dentro de um mesmo emprego são pequenas, para outros departamentos, chefias, ou outra companhia, pois o que você ganha é o valor do que faz, não há como mudar, aumentar.

Resp. 2:

Porque na verdade para você conseguir um aumento tem que ocorrer uma mudança na chefia, nos departamentos ou nas companhias.

Eles determinam (têm) a imagem) de que você tem "aquele" certo salário e não percebem que este pensamento tem que ser maior, para ajudar a própria empresa.

A empresa necessita do empregador, mas este empregado tem sempre que estar melhorando na empresa (sendo que para esta a questão salarial entra na comodidade)

Comentário das respostas: Nas duas respostas podemos perceber a falta de coerência e de coesão. A dificuldade de redação em LM fica evidente, o que denota uma deficiência no domínio do registro de textos escritos. Muito mais que um problema de compreensão de textos em LE, este é um problema de escrita em LM.

Em conversa posterior à prova com os dois alunos que redigiram as respostas acima, antes de dar qualquer feedback sobre a correção, pude constatar que oralmente eles conseguiram responder satisfatoriamente à questão. Isso confirma a proposição acima de que a habilidade de escrita é mesmo uma fonte de variação na correção.

Exemplos de resposta esperada:

Resp. 3:

Porque a empresa na qual você trabalha acha que você está trabalhando por um certo salário e que eles não precisam te pagar a mais pra você continuar fazendo a mesma coisa.

Resp. 4:

Porque eles (os empregadores) compreendem que você está trabalhando por um salário certo, e então eles não deveriam ter que dar de repente a você muito mais para fazer a mesma coisa.

Comentário das respostas esperadas: As respostas acima foram produzidas por outros dois alunos da mesma turma.

Considerando o contexto desta pesquisa (teste de suficiência, envolvendo um grande número de candidatos), percebe-se que o examinador não tem como acessar a compreensão desses leitores, a não ser por meio da resposta dada. Pode-se concluir, então, que o conhecimento que o leitor tem

do registro que deve ser utilizado na formulação das respostas, ou seja, a capacidade discursiva do leitor deve ser considerada uma fonte de variação na correção, em testes discursivos como o proposto neste trabalho. Sendo assim, é interessante ressaltar que esta variável foi considerada na correção das respostas, por ser um dos critérios que possibilitam o nivelamento dos leitores. Não foi, porém, uma variável desenvolvida teoricamente, porque, conforme antes exposto, a análise teórica dessa variável desviaria o foco do objetivo que se quer atingir, inserindo outras tantas variáveis, que precisam ser analisadas adequadamente em algum outro trabalho mais específico sobre o assunto.

3.9.2 Variabilidade relacionada à tarefa (e ao texto)

A segunda fonte de variação apontada por McNamara e citada no início é a opção de escolha de tarefa a ser realizada. No caso desta pesquisa o tipo de tarefa é o mesmo para todos, evitando a variação em função da possível escolha, porém, o tipo de tarefa que o elaborador seleciona para ser realizado no teste também influencia no desempenho dos leitores. Antes, porém, de tratar desse assunto, é interessante observar que existe uma possibilidade de o texto se constituir em uma fonte de variabilidade.

No teste de suficiência proposto, a variação do texto na avaliação também pode exercer alguma influência nos resultados finais do processo avaliativo. Pode-se encontrar um exemplo no processo de avaliação da UFPR no presente momento. Os departamentos da universidade têm autonomia para decidir se querem um teste de suficiência específico para o seu departamento ou não e quem vai ser encarregado de elaborar esse teste. O Departamento de Línguas Estrangeiras Modernas do Curso de Letras prepara os testes para aqueles departamentos que não exigem provas específicas e, neste caso, são elaborados três testes enfocando três grandes áreas de conhecimento: tecnológica, biológica e humanas. Dependendo das diretrizes de cada departamento, o candidato pode ter a liberdade de escolher em qual das três áreas de conhecimento ele quer fazer o teste, ou seja, ele pode escolher entre

três textos diferentes. A implicação disso é que se um profissional da área tecnológica, digamos um engenheiro, quiser fazer mestrado na área de educação, mesmo tendo que trabalhar no mestrado com textos relacionados à área de educação, ele pode escolher fazer o teste da área tecnológica, em função do assunto tratado e da linguagem utilizada no texto ser mais familiar.

Alderson afirma que

> a importância do conhecimento antecedente, cultural, de assunto e tópico em compreensão significa que elaboradores de teste devem estar conscientes de que é bem possível que tal conhecimento influencie escores de teste ou medidas de leitura. Normalmente nós não estamos interessados em medir tal conhecimento em testes de leitura: isto poderia representar uma redução na validade da nossa medida. Uma precaução, então, pode ser selecionar textos em tópicos que são reconhecidos como sendo igualmente familiares ou não familiares a todos os candidatos.[91] (2000,p. 81,tradução nossa).

Apesar do fato de que neste contexto, como apontado acima, não há possibilidade de escolha de tarefa, uma vez que todos os leitores terão que realizar as mesmas tarefas propostas, cabe ao elaborador do teste escolher o tipo de tarefa considerado mais apropriado aos propósitos do teste. É importante salientar que o tipo de tarefa a ser realizado no teste deve ser o mais semelhante possível às tarefas que ele realiza ou deve realizar em situações reais de uso da língua. De acordo com Alderson

> a influência do propósito no processamento e entendimento do texto sugere que avaliadores precisam pensar nas suas questões/tarefas de teste como propósitos de leitura. Quanto mais perto elas chegarem dos propósitos da vida-real, dentro dos limites óbvios de situação de teste, mais probabilidade nós temos de conseguir resultados de teste que vão generalizar e apresentar uma imagem válida daquele tipo de leitura em particular.[92] (2000, p. 82-3, tradução nossa).

---

[91] "The importance of background, cultural, subject and topic knowledge in comprehension means that test designers must be aware that such knowledge may well influence test scores or measures of reading. Normally we are not interested in measuring such knowledge in reading tests: this would represent a reduction in the validity of our measure. One precaution, then, can be to select texts on topics which are known to be equally familiar or unfamiliar to all candidates."

[92] "The influence of purpose on text processing and understanding suggests that testers need to think of their test questions/tasks as reading purposes. The closer they can come to real-life

Em se tratando de avaliações de leitura, “É importante entender que não há nenhum ‘método melhor’ para testar leitura. Nenhum único método de teste pode preencher todos os vários propósitos para os quais nós podemos testar.” (ALDERSON, 2002, p.203, tradução nossa). Alguns tipos de teste são mais utilizados por questões de conveniência e eficiência (são mais rápidos e fáceis de corrigir, por exemplo), mas seria ingenuidade pensar que por serem mais usados esses testes podem ser considerados mais válidos que os outros.

Segundo Bachman e Palmer (1997) várias pesquisas sobre os efeitos dos métodos de testes no desempenho dos testes levam à conclusão de que “é sempre provável que as características das atividades usadas afetem a pontuação dos testes em algum grau, de tal forma que virtualmente não há nenhum teste que dê somente informações sobre a habilidade que nós queremos medir” (p.46, tradução nossa) e, “visto que não podemos eliminar totalmente os efeitos das características da atividade, nós devemos aprender a entendê-los e controlá-los de maneira a assegurar que os testes que vamos usar têm as qualidades que desejamos e são apropriados para os usos para os quais eles são pretendidos.” (p. 46, tradução nossa).

Alderson (2002) apresenta alguns tipos de métodos, ou atividades[93], que podem ser utilizados em avaliações de leitura. Embora não seja o objetivo desta pesquisa classificar tipos de itens de avaliação, e nem se pretenda aprofundar cada uma deles, alguns tipos, que ele e outros estudiosos consideram em suas classificações, serão apresentados a seguir, no intuito de proporcionar um melhor entendimento da discussão sobre o tipo de questão selecionado para o teste da pesquisa, que se fará em seguida.

Cloze test: este tipo de item é construído no próprio texto. Uma ou duas sentenças são mantidas intactas no início e no final do texto, para

---

purposes, within the obvious limits of the testing situation, the more likely we are to get test results that will generalize and present a valid picture of that particular type of reading.”

[93] Em função do emprego de Alderson do termo *método,* neste momento o emprego do termo método e atividade são equivalentes aos termos item (de teste) e questão, devendo ser considerados sinônimos. Dar-se-á preferência, no entanto, aos termos item e questão, para evitar possíveis ambigüidades.

estabelecer maior coesão textual, e um número variável de palavras é retirado do texto, para ser recolocada, ou identificada. Retira-se a enésima palavra do texto, independente de que palavra seja, sendo que 'n' geralmente varia entre cada 5ª e cada 12ª palavra.

Preencher de lacunas (Gap-filling): este tipo diferencia-se do anterior pelo fato de que as palavras retiradas são escolhidas racionalmente e não por um procedimento pseudo-fortuito, isto é, o que importa não é o número de palavras entre uma lacuna e outra, mas a palavra em si. Procura-se deixar um mínimo de 5 a 6 palavras entre cada lacuna, para permitir a contextualização.

Múltipla escolha (Multiple-choice): nestes itens, para cada pergunta são propostas várias alternativas para que se escolha uma ou mais entre elas. Este é um tipo de questão freqüentemente usado em avaliações de leitura.

Relacionar (Matching): item em que se deve fazer a correspondência entre dois grupos de informações, por exemplo, notícias de jornal e suas manchetes.

Ordenar (Ordering): como o próprio nome diz, o objetivo destes itens é ordenar um grupo de palavras, sentenças, parágrafos ou textos, que estão colocados fora de ordem. A ordem pode ser, por exemplo, de importância, cronológica, de seqüência lógica, ou de acordo com o texto.

Verdadeiro ou Falso (Dichotomous): nesses itens, para cada sentença relativa a um texto dado, existem duas alternativas entre as quais se deve escolher: verdadeiro ou falso / certo ou errado / concorda ou discorda do texto. Às vezes há uma terceira alternativa, como: não mencionado / não diz, ou algo assim.

Editar (Editing): este tipo de item consiste de um texto (ou uma parte de) contendo erros, que deverão ser identificados e, possivelmente, corrigidos.

Respostas curtas (Short-answer): esta é uma alternativa semi-objetiva para as questões de múltipla escolha. As perguntas feitas devem ser respondidas de forma breve, com poucas palavras, mas não apenas com Sim/Não ou Verdadeiro/Falso.

Transferência de informação (Information-transfer): o objetivo destes itens é identificar no texto alvo as informações requisitadas e transferi-las, freqüentemente, para uma tabela, um mapa, um esquema, ou para outra

forma qualquer. As informações podem ser constituídas de números, nomes, pequenas frases ou sentenças; e podem ser pontuadas objetiva ou subjetivamente.

Questão aberta (Open-ended): Este tipo de questão não faz parte da lista proposta por Alderson (2002), aqui apresentada e, talvez por isso mesmo, ele não esclareça o que entende por "open-ended questions" (questões abertas), limitando-se a colocá-las em oposição às "closed questions" (questões fechadas). No entanto, considerando a existência de métodos que não poderiam ser classificados em nenhum dos tipos por ele relacionados, há a necessidade de apresentar a visão de outros autores a respeito desta questão.

Davies et al no Dictionary of language testing (2002, p.32) adota as atividades de questão aberta (open-ended question) como sinônimo das chamadas de resposta construída (constructed-response) e resposta extensa (extended response), classificando não a questão em si, mas o tipo de resposta que se espera obter através da questão. De acordo com esses autores este tipo de questão distingue-se das chamadas de escolha forçada (forced-choice) no fato de que no primeiro o leitor deve formular sua própria resposta e neste último o leitor deve escolher uma das opções propostas. Além disso, eles consideram alguns dos métodos (itens) classificados por Alderson como os mencionados acima, formas de respostas construídas, por exemplo: "cloze" e respostas curtas, ou formas de atividades de escolha forçada.

Bachman (1991, p.129) estabelece uma classificação parecida com a de Davies *et al.* Muda somente os termos e não o que se entende por eles, e propõe dois tipos de respostas esperadas: *reposta selecionadas* e *respostas construídas*. Bachman e Palmer (1996, p. 54) dividem os tipos de resposta em três: *resposta selecionada*, em que o indivíduo tem que selecionar uma ou mais respostas, entre as alternativas possíveis; *resposta de produção limitada*, que podem ir de uma palavra a uma sentença e *resposta de produção extensa*, que são livres e podem ter duas ou mais sentenças. Nuttall (2000, p.186) discute as *questões abertas*, as *de múltipla escolha* e as de *verdadeiro ou falso,* como itens pertencentes a um mesmo nível de classificação. Ela afirma que nas questões abertas o aluno pode dar qualquer resposta que considerar

apropriada, mas que essas questões são particularmente aplicadas a perguntas “Q” (Que, Quem, Quando), “Como” e “Por que”.

Entre os estudiosos citados, apenas Nuttall (2000, p.186-7) em sua classificação, apresenta claramente vantagens e desvantagens dos itens chamados de *questões abertas*, que se aplicam às questões elaboradas para os testes desta pesquisa. As duas desvantagens são que essas questões não podem ser avaliadas objetivamente, o que pode dificultar bastante a correção e que nem sempre o leitor é capaz de fazer uso da segunda língua para produzir a resposta, sendo necessária a aceitação do uso da LM na elaboração da resposta. As três vantagens que ela cita são: a relativa facilidade de elaboração das questões, em relação às de múltipla escolha; o fato de que as questões abertas podem ser usadas, praticamente, para qualquer propósito e, principalmente, o fato de elas permitirem que o avaliador tenha um acesso maior à capacidade de compreensão, elaboração e síntese, do leitor.

Vários testes internacionalmente (re)conhecidos utilizam questões objetivas e não discursivas para avaliação de leitura compreensiva. De acordo com Tumolo (2005) o módulo de leitura acadêmica da avaliação do IELTS[94] é composto por três textos seguidos de questões de múltipla escolha, questões de respostas curtas, questões de completar, entre outras de resposta selecionada e de produção limitada e o TOEFL[95] tem como característica avaliar os diferentes níveis de compreensão com o uso de questões de múltipla escolha. Em sua análise desses testes, Tumolo (2005) encontrou em ambos, de modo geral, validade de construto (habilidade de leitura compreensiva). No entanto, apontou a existência de algumas questões irrelevantes ao construto, tais como algumas que testavam conhecimento de vocabulário, ou conhecimentos anteriores sobre o assunto. Alderson (2002, p. 98) afirma que “Bachman et al. (1998) descobriram que quase 70% da variação na dificuldade dos itens no sub-teste de leitura TOEFL poderia ser explicados por aspectos do conteúdo do teste relacionados à gramática e ao conteúdo acadêmico e atual dos itens de leitura”. (tradução nossa). Além disso, ele afirma que em relação ao teste IELTS, ele (ALDERSON, 1993) encontrou “elevada correlação entre

[94] International English Language Testing System

[95] Test Of English as a Foreign Language

um teste de gramática comunicativo e testes de habilidade de leitura acadêmica. Pareceu que algumas vezes o testes de 'gramática' estava mais intimamente relacionado a um teste de leitura acadêmica do que o teste de leitura estava relacionado a outro teste de leitura paralelo." (tradução nossa). Esse tipo de correlação pode levar à noção equivocada de que existe uma equivalência entre o conhecimento gramatical da língua e a habilidade de compreensão leitura.

Embora a gramática, o vocabulário e conhecimento prévio do leitor exerçam grande influência e, em especial os dois primeiros, ocupem um papel importante na compreensão de textos, o propósito dos testes de suficiência para cursos de pós-graduação não é testar estes conhecimentos em si, que devem, em princípio, fazer parte do conhecimento do leitor, para que possam ser utilizados como parte das estratégias de leitura, no decorrer do processo de compreensão do texto, influenciando assim, na qualidade da compreensão. A autora acredita que os testes de proficiência aplicados em larga escala, como IELTS e TOEFL, entre outros, não são os mais adequados para serem utilizados como testes de suficiência para cursos de pós-graduação. A especificidade do propósito destes testes demanda determinadas análises dos resultados que os testes em larga escala como esses acima não possibilitam.

Nesta pesquisa, assume-se que as questões abertas possuem, de fato, as vantagens apresentadas por Nuttall, acima, quais sejam: de serem mais fáceis de elaborar do que as questões fechadas, servirem também ao propósito de avaliar candidatos aos cursos de pós-graduação de mestrado e doutorado e permitirem que o avaliador tenha acesso à capacidade de compreensão, elaboração e síntese, do leitor. Esta última vantagem diz respeito à análise de resultados supra mencionada. Ela é especialmente importante, porque, parte do pressuposto de que um dos requisitos indispensáveis à conclusão desses cursos é a produção de uma monografia (dissertação ou tese). Nessa monografia, além dos objetivos, encaminhamento e resultados da pesquisa, devem constar justificativas e embasamentos teóricos, que serão necessariamente feitos a partir da leitura compreensiva de vários outros textos, tanto em LM quanto em LE. Na opinião desta autora, é fundamental, neste caso, que o avaliador possa ter acesso à capacidade que o leitor tem de

elaborar e sintetizar seu entendimento do texto em LE, pois é para a capacidade de realização desta tarefa em especial que ele estará sendo testado.

Retomando, também, as desvantagens que Nuttall apresenta, vê-se que, de fato, essas questões não podem ser avaliadas objetivamente, o que pode dificultar bastante a correção. Isso, porém pode ser até certo ponto controlado e ter seus efeitos minimizados com a definição de descritores que sirvam de base para a correção, com o estabelecimento de uma escala de classificação para pontuação dos testes e com o treinamento dos corretores[96]. Também é fato, especialmente nos testes em questão, que nem sempre o leitor é capaz de fazer uso da segunda língua para produzir a resposta, sendo necessária a aceitação do uso da LM na elaboração da resposta. Os candidatos ao ingresso em cursos de mestrado e doutorado são das mais diversas áreas de atuação e a experiência docente da autora demonstra que seu conhecimento de inglês pode variar desde o conhecimento básico, obtido no ensino fundamental e de segundo grau, com alguma experiência prática (em geral nas habilidades receptivas) em função das necessidades individuais de estudo, trabalho ou lazer, até o conhecimento avançado e proficiente das quatro habilidades (fala, escrita, leitura e audição). Assim sendo, no caso destes testes de suficiência o uso da LM não representa um problema, mas uma solução, considerando ainda que, no caso das normas seguidas pela UFPR, as respostas dos testes de suficiência **devem** ser dadas em português, conforme Resolução n.º 62/03 – CEPE – Seção VII – Art. 38 – parágrafo 9º - reproduzida a seguir:

"§ 9º O teste de suficiência em língua estrangeira deverá ser respondido na língua portuguesa."

Finalmente, corroborando o argumento proposto de que o domínio da escrita em LE não se faz necessário neste caso, deve-se considerar que as monografias são, **necessariamente**, redigidas em LM, conforme Resolução n.º 62/03 – CEPE, Seção XI - Art. 59, que no parágrafo 3º afirma que:

---

[96] Estes assuntos serão tratados oportunamente, ainda neste capítulo.

"§ 3º - É vedada a apresentação de exemplares finais de dissertação ou tese produzidos em língua estrangeira."

3.9.3 Variabilidade relacionada ao corretor

Com respeito à correção de avaliação, McNamara diz que "julgamentos que valem à pena vão ser inevitavelmente complexos e envolver atos de interpretação por parte do corretor e, portanto, estar sujeito a discordâncias". (1996, p.117) Podemos, assim, dizer que a análise das respostas, no caso das questões discursivas, é necessariamente subjetiva; e mais do que classificar cada questão como 'certa' ou 'errada', os examinadores precisam fazer julgamentos sobre o quão correta/incorreta ou abrangente é a resposta.

> Introduzir o corretor no processo de avaliação é ao mesmo tempo, necessário e problemático. É problemático porque correções são necessariamente subjetivas. Outra forma de dizer isso é que a correção dada a um candidato é um reflexo, não apenas da qualidade do desempenho, mas das qualidades como corretor da pessoa que corrigiu."[97] (MCNAMARA, 2000, p.37, tradução nossa).

Em outras palavras, o que McNamara quer dizer é que não há como evitar ou eliminar completamente algum tipo de variação entre corretores. No entanto, o fato de ela existir não significa que não deva ter um limite e, menos ainda, que deva ser aceita incondicionalmente.

De acordo com Alderson, Clapham e Wall (1995) o trabalho do examinador, mais do que apenas dizer se uma resposta está certa ou errada, "é avaliar quão bem um candidato completa uma dada tarefa, e para fazer isso eles precisam de uma escala de classificação" (tradução nossa). Essa escala

---

[97] "Introducing the rater into the assessment process is both necessary and problematic. It is problematic because ratings are necessarily subjective. Another way of saying this is that the rating given to a candidate is a reflection, not only of the quality of the performance, but of the qualities as a rater of the person who has judged it."

pode ser holística, e julgar o desempenho do candidato como um todo, ou analítica, e julgar aspectos específicos (detalhes) desse desempenho, dependendo da intenção do elaborador do teste. Os autores afirmam que em ambas as análises é imprescindível que sejam estabelecidos descritores[98] para cada componente avaliado.

Como mencionado anteriormente, em se tratando de avaliações como estas de suficiência, os julgamentos feitos trarão conseqüências diretas a quem estiver sendo julgado. Este fato, necessariamente, envolve questões de justiça em relação aos procedimentos utilizados pelos juízes em questão: os corretores dos testes.

Entende-se que algum grau de divergência na correção de testes que envolvem itens discursivos e corretores é previsível e aceitável, como afirmou McNamara. Porém, as diferenças devem ser minimizadas o máximo possível, para que as decisões tomadas a partir dos resultados da correção colaborem na garantia da validade de todo o processo de avaliação.

A correção tem influência direta na validação dos testes, porque em todo processo de avaliação que envolva julgamento humano, seja de um ou de mais corretores, haverá sempre algum tipo de divergência em função da subjetividade característica dos julgamentos humanos. Vejamos a descrição de dois exemplos:

Situação: 01teste com 10 questões discursivas. Quantidade de testes para correção: 100 testes	
Um único corretor: neste caso as possíveis causas de variabilidade na correção podem ser, entre outras:	• O corretor não consegue corrigir tudo em um único dia, mesmo que ele corrija o dia todo. • Ao longo do dia o cansaço pode ser fonte de variação e o corretor pode, conseqüentemente, ficar mais exigente ou mais leniente. • A variação que se repete ao longo do dia, pode repetir-se de um dia para outro, também, porque

[98] O estabelecimento de descritores, bem como as escalas de avaliação são assuntos tratados no item “Definição do Construto”.

	quanto mais cansado, mais exigente ou leniente o corretor pode ficar. • Outra variação pode ocorrer no humor do corretor, em diferentes dias, interferindo no seu padrão de correção mais ou menos. • Dependendo de quão bem estabelecidos estejam os critérios de correção, as variações na correção de uma mesma questão podem surgir, em maior ou menor grau, ao longo das horas e dos dias.
Mais de um corretor: neste caso, além das variáveis individuais, mencionadas acima, podemos ter:	• Diferenças no (não) estabelecimento de critérios de correção. • Diferenças quanto à rigidez de obediência aos critérios. • Diferenças quanto à compreensão dos critérios. • Maior ou menor influência das variações de humor de cada corretor; etc.

TABELA 3 – EXEMPLOS DE CORREÇÃO.
FONTE: AUTOR (2008)

Como se pode observar por estes exemplos, o estabelecimento de critérios e a chegada a um consenso, entre os corretores, sobre como deva ser o processo de correção é de extrema importância, ou todas essas influências e diferenças podem tornar o processo de avaliação inválido, na medida em que, para um corretor um candidato pode obter um escore acima do limite mínimo estabelecido, sendo assim aprovado na seleção; enquanto para outro corretor, este mesmo candidato pode obter escore abaixo do limite mínimo, não sendo, portanto, aprovado. Dependendo do número de diferenças como esta, o resultado da avaliação pode sofrer alterações consideráveis em termos da qualidade dos candidatos aprovados.

De acordo com McNamara (2000, p.36) as avaliações mediadas por corretores estão cada vez mais presentes no processo de ensino e aprendizagem de línguas, porque o processo comunicativo utilizado está concentrando-se mais no desempenho comunicativo contextualizado do

aprendiz e, conseqüentemente, julgar o impacto dessa comunicação passou a ser o foco da avaliação do uso da língua alvo. Para esse autor, as "diferenças entre juízes podem ser entendidas em termos de *severidade*[99] (ou indulgência) global por um lado, e aleatoriedade do erro (*erro*) por outro". (1996, p. 122, tradução nossa).

Mcnamara (1996, p. 123) apresenta algumas formas nas quais os corretores podem diferir, que ele considera importantes, quais sejam:

a) Eles podem diferir em termos de indulgência global. Ou seja, diferir em termos de características gerais de cada corretor;
b) Eles podem diferir no que o autor chama de "interação corretor-item", ou "interação corretor-candidato". No primeiro caso, os corretores podem ser mais rigorosos ou indulgentes na correção de algum(ns) item(ns) e de outro(s) menos. No segundo caso, os corretores podem ser mais rigorosos ou indulgentes com um grupo de candidatos do que com outro. Entretanto, em geral, segundo o autor, o grau de severidade ou indulgencia dos corretores é consistente nos dois casos, ou seja, eles serão consistentemente severos ou indulgentes com um mesmo grupo de candidatos ou na correção de um mesmo item para todos os candidatos.
c) Eles podem diferir na maneira de interpretar a escala de classificação utilizada. Aparentemente, as escalas de avaliação possuem intervalos iguais[100], visualmente algo como:

0 1 2 3 4 5

No entanto, a interpretação de um corretor desta escala pode variar, pois para um corretor a distância entre um candidato cuja habilidade é classificada na categoria 0 e outro na categoria 1 pode ser menor que para outro corretor, e assim sucessivamente com relação a todas as categorias da

[99] Em nota McNamara sugere o termo *características do corretor* em lugar de *severidade*.

[100] Os intervalos das escalas são "aparentemente" iguais, porque estas são escalas ordinais e não intervalares. Ver definição de escalas no item 2.5.4.2 acima - "A natureza das medidas e as escalas".

escala. Um exemplo de duas interpretações diferentes poderia ser da seguinte forma:

Essa diferença de interpretação representada pelas duas escalas acima significa que, por exemplo, para o primeiro corretor a diferença entre candidatos da categoria 1 e da categoria 0 é bem pequena, entre 1 e 2 a diferença é bem maior, ou seja, para que um candidato passe da categoria 1 para a 2 o seu desempenho tem que dar um “salto” bem maior do que para passar da categoria 0 para a 1. Para o segundo corretor o “salto” entre as categorias 1 e 2 é bem menor que entre 0 e 1.

Supondo que um candidato esteja na interseção de duas categorias, ou seja, se ele não tiver a mesma habilidades de outros classificados da categoria 2, mas, por outro lado, tiver uma habilidade superior à maioria dos classificados na categoria 1, o corretor terá necessariamente que fazer uma escolha por uma das duas categorias. Como a divisão das categorias é estabelecida em intervalos diferentes por cada corretor, este candidato poderia ser classificado diferentemente por cada corretor, dependendo da interpretação que cada um fizer da escala original (com intervalos aparentemente iguais).

Eles podem, por fim, diferir em termos da extensão de erros aleatórios associados à sua correção. Sua maior ou menor consistência em relação à correção que outros corretores fazem, dos mesmos candidatos. O corretor pode ser irregular e não seguir um padrão de severidade ou indulgência em sua correção como um todo; a classificação que ele faz dos candidatos pode não permitir uma relação consistente entre a sua e a classificação dos mesmos candidatos por outro corretor.

A falta de consistência apresentada no item 4 acima, de acordo com o mesmo autor, é difícil de ser eliminada ou compensada, por isso, corretores desse tipo devem ser treinados novamente ou eliminados, caso o novo treinamento não resolva o problema. Nos outros casos, o estabelecimento de critérios, o treinamento dos corretores e programas estatísticos de análise de

dados podem fazer a equalização dos escores médios dados pelos corretores, minimizando a influência dessas fontes de variabilidade nos resultados das avaliações.

3.9.4 Treinamento de Corretores

Em função da variabilidade que pode existir entre os corretores e das conseqüências por elas acarretadas, é natural e mesmo óbvia a afirmação de McNamara (1996) de que seria muito mais simples se não houvesse tanta diferença entre os corretores, principalmente no que diz respeito aos erros aleatórios. No entanto, o fato é que as diferenças existem e uma das formas de se tentar reduzir o efeito delas no processo de avaliação é o treinamento de corretores.

Embora, de acordo com McNamara os efeitos do treinamento de corretores não venham sendo estudado, como talvez devessem, algumas pesquisas demonstram que:

> 1. Treinamento de corretores é bem sucedido em tornar corretores mais auto-coerentes. Isto é, o efeito principal do treinamento é reduzir o erro aleatório nos julgamentos do corretor.
> 2. Treinamento de corretores pode reduzir, mas de forma alguma eliminar a extensão da variabilidade do corretor em termos de severidade global. Em especial, diferenças extremas são reduzidas [...] mas diferenças significativas e substanciais entre corretores persiste. Lunz e Stahl (1990) argumentam que juízes empregam percepções únicas que não são facilmente alteradas por treinamento."[101] (1996, p. 126, tradução nossa).

---

[101] "1. Rater training is successful in making raters more self-consistent. That is, the main effect of training is to reduce the random error in rater judgments.

2. Rater training can reduce but by no means eliminate the extent of rater variability in terms of overall severity. In particular, extreme differences are reduced […] but significant and substantial differences between raters persist. Lunz and Stahl (1990) argue that judges employ unique perceptions which are not easily altered by training."

Alderson, Clapham e Wall (1995, p. 105) defendem que “o treinamento de corretores é um componente crucial de qualquer programa de teste, na medida em que, se a pontuação do teste não é válida e confiável então todo o outro trabalho realizado antes para elaborar um instrumento ‘de qualidade’ terá sido uma perda de tempo.” (tradução nossa).

Segundo eles na pontuação subjetiva de testes os corretores fazem julgamentos mais complexos do que as decisões de “certo e errado”. Os corretores avaliam a qualidade de desempenho do leitor em cada tarefa executada. Para isso, utilizam uma escala de classificação que pode ser holística ou analítica, conforme mencionado no início do item 2.12.3 acima. Quando os corretores utilizam uma escala holística, como é o caso nesta pesquisa, eles ficam atentos ao desempenho global do leitor e não se concentram em componentes específicos da linguagem. No caso da leitura, isso significa, por exemplo, que o corretor não vai verificar se o leitor usa exatamente as mesmas palavras do texto na elaboração das respostas, isto é, o leitor tem uma flexibilidade maior no emprego de sinônimos e paráfrases dos termos e idéias apresentadas no texto, desde que o significado seja preservado e dentro de certos limites estabelecidos pelo próprio texto.

Os autores acima (1995, p. 110-13) fazem algumas recomendações em relação ao treinamento de corretores de avaliações de escrita, que também se aplicam às avaliações de leitura e que serão apresentadas de forma adaptada em forma de itens a seguir.

a) o elaborador do teste deve ser idealmente o elaborador da escala de classificação que será utilizada pelos corretores;
b) não é recomendável a utilização de escalas contendo apenas valores numéricos ou conceitos como “ótimo”, “bom”, “regular”, etc, pois a interpretação dos valores e conceitos pode variar entre os corretores;
c) utilizar escalas de no máximo sete (07) níveis de pontuação, pois consideram difícil estabelecer distinções mais precisas para um número de descritores maior que este;
d) deve haver entre os corretores um corretor ‘chefe’ (C);

e) após a aplicação do teste o corretor C deve ler rapidamente vários testes para levantar os tipos de respostas e problemas que elas podem apresentar;
f) O corretor C deve, então, extrair algumas respostas que possam ser consideradas 'adequadas' e 'inadequadas', bem como problemas que geralmente os corretores enfrentam, mas que não constam nos descritores, como letra ilegível;
g) usar a escala de classificação na amostra de respostas selecionada, para estabelecer os padrões de correção. Isso deve ser feito, preferencialmente, pelo corretor C e mais alguns (poucos) corretores selecionados por ele, no caso de haver um grande número de corretores;
h) esse pequeno comitê de corretores deve, então, comparar suas pontuações, discutir as diferenças de opinião e estabelecer, de forma consensual, uma pontuação para as amostras, a partir da qual, eles melhorarão a escala de classificação, se necessário, para que ela fique mais compreensível e fácil de usar;
i) quando houver um grande número de corretores, deve ser realizada uma reunião de padronização da correção, na qual o comitê descrito acima deverá treinar os demais corretores;
j) antes da reunião de treinamento, todos os corretores devem receber as amostras dos testes, para usar neles a escala de classificação. Assim, no momento do treinamento, os corretores não serão influenciados pela opinião do comitê. Serão levantados possíveis novos problemas encontrados pelos corretores e trabalhada a padronização das pontuações destes com a do comitê;
k) antes que a correção comece, o corretor C deve incorporar as possíveis alterações acordadas na reunião de treinamento e providenciar cópias desta escala para todos os corretores.

Alderson, Clapham e Wall defendem a idéia de que todos os corretores devem passar por esse tipo de treinamento a cada correção de um novo teste, mesmo aqueles corretores mais experientes, para evitar que algum comportamento peculiar de correção individual interfira no processo.

4 DESENVOLVIMENTO E APLICAÇÃO DOS TESTES

Na elaboração de um teste há várias questões importantes a serem consideradas. Uma delas é o quê e como queremos avaliar. Alderson (2002, p.117) afirma que em um teste de leitura, o que se quer testar realmente é quão bem e com que grau de eficiência alguém é capaz de ler um texto, qualquer que ele seja, para um propósito previamente definido. Afirma ainda, que o método utilizado no teste é uma fonte de influência em potencial na medição. Isso significa que uma vez que "o que nós queremos, na verdade, medir em um teste é o resultado de uma interação entre habilidade e método de teste, o uso de métodos de teste apropriados pode melhorar a validade das nossas inferências". (ALDERSON, 2002, p.117). No final de uma avaliação, o que nos interessa é poder fazer generalizações a partir dos resultados obtidos, relacionando o desempenho obtido nos testes ao desempenho nas tarefas a serem realizadas na vida real[102]. Como afirma Alderson (ibidem) "Nós não estamos interessados em saber quão bem um leitor pode fazer nosso teste, nós queremos saber alguma coisa sobre sua habilidade de leitura ou procedimento de leitura além da situação de teste – que é freqüentemente referido como a generalização dos resultados do nosso teste." Uma forma de lidar com essa generalização é através do construto do teste.

Neste capítulo serão apresentados: a metodologia utilizada no desenvolvimento desta pesquisa; o desenvolvimento dos testes aplicados, em relação à definição do construto, à escolha do texto e a definição dos itens de avaliação, i.e., das questões; a correção dos testes-piloto; a escolha de um dos testes e sua re-elaboração para o teste final.

[102] Entende-se vida real, aqui, em oposição à situação de avaliação que, embora parte da vida real, é realizada em um contexto e ambiente especialmente determinados para este fim.

4.1 METODOLOGIA DE PESQUISA

Segundo Nunan (2003, p.3), tradicionalmente, é feita uma distinção binária entre pesquisa *quantitativa e qualitativa*, sendo a primeira uma pesquisa controlada, objetiva, generalizável e que assume a existência de fatos que são independentes e externos ao observador e ao pesquisador, e a segunda é aquela em que o conhecimento é relativo, em que há um elemento subjetivo para todo conhecimento e pesquisa e cujos estudos holísticos e não generalizáveis também são justificáveis. Entretanto, alguns dos autores citados por Nunan (2003, p.4-10) afirmam que esta distinção é simplista e ingênua. O primeiro deles é Grotjahn[103] (1987), segundo o qual esta distinção binária qualitativa/quantitativa é uma simplificação excessiva. Ele propõe que os estudos atuais sobre pesquisa estão levando em consideração "o método de coleta de dados (se os dados devem ser coletados experimentalmente ou não-experimentalmente); o tipo de dados produzidos pela investigação (qualitativos ou quantitativos); e o tipo de análise dos dados conduzida (se estatística ou interpretativa)"[104], sendo que a combinação dessas variáveis geram paradigmas (assim denominados por Grotjahn), que compõem os tipos possíveis de pesquisas.

O segundo autor é Van Lier[105], que segundo Nunan apresenta dois parâmetros de pesquisa em lingüística aplicada, quais sejam: intervenção/não intervenção e altamente seletivo/não seletivo. Estes dois parâmetros geram quatro espaços semânticos, que ele denomina de:

a) espaço de "controlar" (controlling space) – são aqueles estudos em que existe um alto grau de intervenção e de seleção, por parte do

---

[103] **Grotjahn, R.** On the methodological basis of introspective methods. In C. Faerch and G. Kasper (eds.), *Introspection in Second Language Research.* Clevedon Avon, England: Multilingual Matters, 1987.

[104]Tradução da nossa.

[105] **van Lier, L.** Ethnography: Bandaid, bandwagon, or contraband. In C. Brumfit and R. Mitchell (eds.), *Research in the Language Classroom.* London: Modern English Publications, 1990 – p.34.

pesquisador, que se concentra em um número limitado de variáveis, as quais procura controlar de alguma forma;

b) espaço de "avaliar" (*measuring space*) - são as pesquisas que apresentam um alto grau de seleção e baixo grau de intervenção, por parte do pesquisador, cuja intenção é observar sem controlar as variáveis que escolheu para estudar;
c) espaço de "perguntar/fazer" (*asking/doing space*) – são aquelas pesquisas com alto grau de intervenção e baixo grau de seleção, em que se procura fazer uma investigação para identificar algum problema, por exemplo;
d) espaço de "observar" (*watching space*) – são aqueles estudos em que o grau de intervenção e de seleção são baixos, como quando, por exemplo, o pesquisador pretende apenas descrever e interpretar um ambiente ou situação, sem fazer a seleção de alguma variável específica ou exercer qualquer tipo de controle.

Nunan e, de acordo com ele, o próprio Van Lier consideram esta divisão uma simplificação do que acontece nos processos de pesquisa, porque a pesquisa pode transitar entre estes quatro espaços, dependendo da direção que o pesquisador tomar no decorrer do seu estudo.

Brown[106] (1988) é o terceiro autor citado por Nunan e apresenta um sistema de análise de tipos de pesquisa que se divide em primário e secundário. A pesquisa do tipo primário deriva de fontes de informação primárias, tais como a análise de resultados de testes e pressupõe a existência da pesquisa secundária como pré-requisito necessário. A pesquisa secundária deriva de fontes secundárias tais como um livro sobre a análise de resultados de testes e compreende uma revisão bibliográfica em determinada área, e a síntese da pesquisa realizada por outros pesquisadores. A pesquisa primária subdivide-se em estudo de caso e estatística, que por sua vez, subdivide-se em levantamento, que investiga atitudes, opiniões ou características (freqüentemente através de questionários) e estudos experimentais, que controla as condições sob as quais são realizadas as observações.

---

[106]**Brown, J. D.** *Understanding Research in Second Language Learning: A Teacher's Guide to Statistics and Research Design.* New York: Cambridge University Press, 1988.

Nunan (2003. p.10) defende que, apesar da distinção entre pesquisa qualitativa e quantitativa ser simplista e ingênua, como afirmam os autores citados, essa distinção ainda persiste e que isso se deve, em parte, porque “as duas abordagens representam maneiras diferentes de pensar sobre e entender o mundo entorno de nós.” Além disso, existe um debate filosófico, sobre a natureza do conhecimento e sobre como são feitas as afirmações a respeito do mundo, que subjaz ao desenvolvimento de diferentes tradições de pesquisa e de métodos.

Considerados os diferentes pontos de vista sobre os métodos de pesquisa e o contexto específico em que esta pesquisa está inserida, a autora defende o ponto de vista de que não seria necessário, ou mesmo possível a utilização de uma teoria única na classificação do tipo de pesquisa realizada.

Assim sendo, quanto à forma de abordagem, esta pesquisa se serviu de duas abordagens diferentes, em momentos diferentes. Num primeiro momento, foi realizada uma pesquisa **qualitativa**, para que pudesse ser estabelecido o construto a ser avaliado, para escolher os textos a serem utilizados no teste aplicado e para embasar a elaboração, correção e análise teórica deste teste. Para isso foi realizada uma revisão bibliográfica sobre avaliação, leitura e tipos de texto, construtos, correção de teste e sobre método de análise de teste, estabelecendo uma comparação entre a TCT e TRI.

Em um segundo momento, foi feita uma análise **quantitativa** dos dados obtidos com a elaboração de quatro testes-piloto e um teste final de suficiência em leitura para fins acadêmicos, seguida da aplicação e correção, com base na revisão bibliográfica e na TRI e com o uso do aplicativo de análise estatística de dados, o Modelo Rasch.

Os sujeitos de pesquisa dos testes-piloto foram:

a. alunos de cursos de graduação e de pós-graduação - especialização (UTFPR);
b. alunos de inglês instrumental e candidatos a cursos de pós-graduação - mestrado ou doutorado em diferentes áreas de estudo (CELIN – UTFPR/PG).

Os sujeitos de pesquisa do teste-final foram:

a) alunos (universitários) dos 2 últimos períodos do curso de graduação de letras (UFPR – TUIUTI – UNIBRASIL);
b) alunos (universitários) dos 2 últimos períodos do curso de Secretariado Executivo Trilíngüe (FACINTER);
c) alunos do curso de leitura instrumental (UFPR – CELIN), candidatos a cursos de pós-graduação, independentemente da área de estudo;
d) alunos de cursos de pós-graduação - especialização – e alunos (universitários) de cursos de graduação da área tecnológica (UTFPR).

Essa escolha possibilitou que se tivesse um público com características similares ou iguais as do público alvo, i.e, candidatos a cursos de pós-graduação – mestrado e/ou doutorado.

Foram selecionados para os testes textos da área de humanas. Esta escolha deveu-se ao fato de grande parte dos assuntos pertinentes a ela poderem ser utilizados na elaboração de testes para outras áreas, coisa que, em geral, as ciências biológicas e exatas não possibilitam tão facilmente. Como um exemplo dessa flexibilidade dos textos classificados como pertencentes à área de humanas, vê-se que um texto sobre depressão, que à primeira vista pertence à área biológica, pode ser escrito sob o ponto de vista das ciências sociais, enfocando as implicações do aumento da incidência desse problema em um determinado grupo social e analisando as possíveis conseqüências desse aumento na sociedade como um todo. Ainda que este texto traga algumas informações técnicas, específicas da área médica, sua classificação não será alterada, uma vez que o assunto é tratado sob a ótica de uma das subdivisões das ciências humanas.

Quanto aos procedimentos técnicos, esta foi uma pesquisa **experimental**, porque as variáveis que poderiam influenciar o objeto de estudo – a avaliação de leitura em inglês como LE – foram consideradas, e formas apropriadas de controle dessas variáveis foram utilizadas. A correção dos testes aplicados foi realizada por professores de inglês, com formação acadêmica em Letras, que foram orientados a corrigir os testes de acordo com o padrão de correção escolhido (detalhado na definição de construto adiante) e a seguir os procedimentos que foram estabelecidos com base na literatura

estudada e apresentada anteriormente no item que trata do treinamento de corretores. Os testes-piloto (total de 94 testes) contaram com dois (02) corretores, sendo que ambos corrigiram todas as provas, ou seja, cada prova foi corrigida duas vezes, uma por cada um dos corretores. O teste final (total de 120 testes) contou com quatro (04) corretores, que corrigiram grupos diferentes de teste, sendo que cada teste também foi corrigido por dois corretores diferentes. Os 120 testes foram divididos em quatro grupos de forma que cada corretor ficou responsável pela correção de 30 testes. Além dos seus 30 testes, cada corretor ainda corrigiu 10 testes relativos aos grupos de cada um dos outros três corretores, conforme tabela abaixo.

Grupos de teste / Corretor	A (nº de testes corrigidos)	B (nº de testes corrigidos)	C (nº de testes corrigidos)	D (nº de testes corrigidos)	TOTAL
1	30	10	10	10	60
2	10	30	10	10	60
3	10	10	30	10	60
4	10	10	10	30	60

TABELA 4- TESTES APLICADOS.
FONTE: AUTOR (2008)

Esse procedimento visou reproduzir a situação real de correção dos testes de suficiência cujo número de candidatos é grande e exige o trabalho de mais de um corretor. Essa similaridade é uma forma de garantir maior validade à correção na pesquisa. Além disso, com a participação as autora da pesquisa na correção, a existência de outro(s) corretor(es) garantiu que os resultados do teste fossem o mais independente possível do avaliador (no caso, a autora).

A elaboração da escala de classificação e o treinamento dos corretores foram realizados pela autora, com vistas a padronizar a correção e obter análises de resultados mais homogêneos, além de dar maior confiabilidade à análise da correção dos testes e à análise dos itens dos testes, que deveria ser (e foi) feita com base na correção e na revisão bibliográfica sobre o assunto.

Esta pesquisa foi **descritiva** quanto aos objetivos, pois objetivou descrever, a partir da revisão bibliográfica, características essenciais

concernentes ao processo de avaliação como um todo, bem como dos sujeitos a serem avaliados. Além disso, a análise dos resultados foi feita com base nos gráficos gerados pelo aplicativo utilizado (Modelo Rasch), por meio de uma descrição e discussão das informações mais relevantes presentes nesses gráficos e em outras informações geradas pelo aplicativo, também apresentadas.

Finalmente, quanto à sua natureza, esta foi uma **pesquisa aplicada**, que objetivou compilar alguns conhecimentos imprescindíveis na elaboração de avaliações de leitura em LE, para que pudessem ser aplicados na elaboração de instrumentos de avaliação para fins acadêmicos em geral e para elaborar e aperfeiçoar um teste de suficiência que pudesse servir de base para elaboração de outros, viabilizando a construção futura de uma escala de medição para avaliações de suficiência em inglês como LE, para cursos de mestrado e doutorado.

4.2 DESENVOLVIMENTO DO TESTE

Para os propósitos desta pesquisa definimos o tipo de teste aplicado como sendo um “teste de Proficiência”, de acordo com a classificação de McNamara, mais precisamente de “Seleção”, segundo a classificação de Bachman e Palmer [107]. O contexto considerado foi o ingresso em cursos de pós-graduação de uma universidade (mestrado ou doutorado).

Como mencionado anteriormente, a variável determinada para o estudo foi a compreensão de leitura. Para que fosse possível a definição do construto a ser avaliado foi preciso definir o modelo de leitura fluente considerado ideal, isto é, o nível de leitura ideal esperado no desempenho dos candidatos aos cursos de mestrado e doutorado. Também foi preciso estabelecer as características fundamentais dos textos considerados relevantes para o caso, ou seja, escolher com que tipo(s) de texto(s) os candidatos deveriam saber trabalhar durante o período dos cursos em questão

[107] Ver a classificação apresentada na seção 3.2.1.1

e, finalmente, selecionar o tipo de teste e itens mais apropriados nesse contexto.

No processo de seleção focado o objetivo prático, geralmente, é saber se o nível de representação mental que o candidato é capaz de fazer é suficiente para que ele possa fazer uso efetivo e eficaz das informações que lhe serão necessárias no decorrer do curso. Para isso, na definição do construto, precisamos levar em conta qual é o nível de representação mínimo considerado necessário para o caso. De certa forma, isso significa o mesmo que estabelecer uma linha de corte que defina o conhecimento que se considera suficiente ou insuficiente. Essa linha de corte vai variar de acordo com os objetivos e expectativas de cada instituição. Portanto, neste trabalho, foram feitas algumas escolhas com base nas diretrizes da UFPR e nos padrões exigidos no processo de seleção em vigor, no momento da pesquisa.

4.2.1 Definição do Construto

A definição do construto do teste foi baseada no levantamento bibliográfico realizado sobre o assunto, apresentada no capítulo 2, item 2.5.3. Entendemos que esta definição, como explicitado no item supracitado, é apenas uma das possíveis abstrações que poderíamos definir para os propósitos deste teste. Todas as escolhas foram feitas levando em consideração as adequações institucionais (UFPR) e a adequação da medição e análise pretendidas (uso do aplicativo – Modelo Rasch), no intuito de preencher os requisitos necessários para a validade do teste.

Retomando o conceito de construto[108], o "Dictionary of language testing" o define como sendo um traço latente, isto é uma variável não observável; um processo psicológico que se quer medir. Conforme Wright (1979, p. 2) a definição dessa variável que se quer medir "começa como a idéia geral do que nós queremos medir. Esta idéia geral ganha significado pela escrita dos itens do teste que visam extrair sinais da variável pretendida no

[108] Ver item 3.2.2

comportamento das pessoas. Estes itens do teste tornam-se a definição operacional da variável." De acordo com ele ainda, conforme visto anteriormente[109], os itens de um teste devem pertencer a uma mesma linha de investigação – que vai exigir níveis de capacidade do mais baixo para o mais alto – para que seja definida a variável a ser investigada no teste. A linha de investigação deste teste foi traçada pelas questões elaboradas, com itens de dificuldade variada, de forma que o candidato pudesse demonstrar o seu grau de compreensão das principais questões apresentadas em gêneros de textos científicos, em especial textos formativos para jornais e revistas especializados (artigos e relatórios de pesquisa, revisões críticas e formulações teóricas) ou textos para popularização da informação científica (resumos de relatórios de pesquisa, artigos de reportagem e comentários críticos). A partir disso, estabeleceu-se que no teste de suficiência em questão:

a) a variável geral medida foi a ***compreensão de leitura***;
b) a definição geral do construto de leitura do teste foi feita com base na escala de compreensão geral de leitura do *Common European Framework* e levando em consideração o propósito dos testes de suficiência. Foi estabelecido como nível geral de compreensão esperado do candidato a mestrado e doutorado o seguinte:
   — capacidade de ler com um grau razoável de independência diferentes gêneros de texto, especialmente os científicos, produzidos pelo e para o meio acadêmico, com a utilização seletiva das estratégias e dos recursos apropriados para alcançar um nível satisfatório[110] de compreensão do texto.
c) os traços específicos observados foram:
   — capacidade de trabalhar com um grande volume de informações;
   — capacidade de entender as idéias gerais apresentadas em textos complexos, que se espera encontrar na vida acadêmica de um mestrando e/ou doutorando;

---

[109] Idem.

[110] O que se entende por nível "satisfatório" de compreensão está explicitado nos descritores dos traços observados.

— capacidade de usar as estratégias de leitura apropriadas para selecionar as informações relevantes (solicitadas) em textos longos;
— capacidade de decidir quando um estudo mais detalhado do texto é necessário para a elaboração da tarefa proposta;
— capacidade de apresentar as informações solicitadas de forma coerente e coesa, sempre que necessário, demonstrando a compreensão do texto lido.

d) a operacionalização do construto se deu pela elaboração de questões, em cuja resposta o leitor pudesse apresentar sua:
— capacidade de encontrar as informações desejadas em textos relativamente longos (sem edições ou cortes) e capacidade de demonstrar a compreensão do assunto por meio de respostas inteligíveis à questões discursivas.

e) os tipos de questões elaborados foram:
— questões implícitas textualmente: isso significa que as respostas das questões são todas encontradas no texto, devem ser retiradas do texto e podem ser uma tradução do texto (quando possível), ou uma elaboração a partir das informações encontradas no texto.
— maior número de questões de compreensão global/geral e menor número de compreensão local/específica.

4.2.2 Critérios de Correção

Para McNamara “uma questão cada vez mais importante na validação de avaliações de desempenho é como os critérios relevantes para acessar o desempenho serão decididos.” (2000, p.37) Em outras palavras, como, baseados em quê devemos definir qual ou quais critérios são, de fato, mais relevantes para acessar determinado tipo de desempenho em um teste?

A idéia é que se os critérios de correção forem estabelecidos de forma clara, o treinamento do corretor será mais adequado e a subjetividade da correção poderá ser minimizada, aproximando o processo de avaliação da objetividade[111].

Bachman e Palmer (1996, p. 212) afirmam que uma das vantagens das escalas construídas com critérios é que elas permitem que se façam inferências sobre qual é a habilidade de linguagem um examinando tem, e não simplesmente quão bem é o seu desempenho em relação a outros indivíduos. De acordo com Alderson (2000, p. 151) os critérios de correção "são supostamente baseados em informações do texto" e as respostas dadas são julgadas de acordo com a extensão de correspondência que a interpretação do examinando alcança em relação às do elaborador do teste, que subentendem a noção de resposta "correta".

A autora entende que a definição do que deva ser considerada uma "resposta correta" é subjetiva e passível de questionamento. Entretanto, não há como eliminar totalmente esta subjetividade, uma vez que cada diferente teoria sobre leitura compreensiva pode defender diferentes aspectos do que uma resposta deva incluir para ser considerada "correta". Dependendo da teoria de leitura subjacente e da interpretação que o elaborador do teste faz dessa teoria (fator também subjetivo que, embora possa ser minimizado, não pode ser eliminado) a resposta pode ser mais ou menos abrangente e fiel ao texto, e os critérios de correção poderão ser mais ou menos rigorosos, incluindo mais ou menos detalhes específicos para o controle de correção de cada resposta.

Dado o fato de que esta pesquisa não adotou nenhuma teoria específica e única de leitura, mas foi embasada em vários conceitos apresentados no levantamento bibliográfico acima, os critérios de correção dos testes elaborados foram, da mesma forma, embasados nos conceitos adotados para a pesquisa, sem privilegiar um ou outro conceito em particular. Esses critérios serão apresentados no item 4.3, que trata da correção dos testes.

---

[111] A discussão sobre objetividade nos testes foi feita no item 3.7, que trata da adequação da avaliação quanto à medida.

4.2.3 Unidimensionalidade do teste

A autora entende que o fato deste teste considerar a capacidade de expressão como um dos traços a serem observados, não compromete a sua característica de unidimensionalidade, pois as respostas são redigidas na LM do candidato, que nessas condições (grau de escolaridade mínimo = universitário completo) a autora assume ter a capacidade de expressão lingüística mínima desejável para a realização das tarefas propostas no teste, uma vez que esta capacidade também está prevista para a escrita da monografia de conclusão de curso. Isto significa que o candidato deve ser capaz de redigir uma resposta, que contenha as informações desejadas e que seja sintática e semanticamente coerente. Sendo assim, a capacidade de redação não foi considerada como uma segunda variável presente no teste, mas como um traço observável, que possibilitou a observação e análise dos outros traços em questão.

4.2.4 Escolha de Textos

No processo de desenvolvimento de um teste de leitura, o primeiro passo é escolher o texto a ser utilizado no teste. Em se tratando este de um teste de suficiência em LE, para fins acadêmicos, optou-se por selecionar gêneros de texto que são recorrentes no meio acadêmico, ao longo dos estudos e pesquisas de mestrado e doutorado. Esta escolha pareceu ser a mais lógica pelo fato de que o teste de suficiência visa assegurar que o candidato tenha domínio e autonomia suficientes de leitura em uma LE, para possibilitar o acesso mais fácil e imediato a materiais e pesquisas nessa língua, que sejam relevantes e até mesmo fundamentais para sua dissertação ou tese.

Seguindo essa linha de pensamento, a primeira opção de gênero com a qual trabalhamos foi o resumo (abstract), que será tratado a seguir. Como segunda opção foram considerados textos científicos, como relatos de

pesquisa, resultados de pesquisa, artigos, etc, e os de divulgação científica, como reportagens e artigos de jornais e revistas especializadas de diversas áreas.

4.2.4.1 Resumos de teses e dissertações

Como dito acima, o resumo (abstract) foi o primeiro gênero de texto considerado para a utilização no teste de leitura. O uso desses textos seria justificado pelo fato de que alunos de mestrado e doutorado precisam necessariamente ser capazes de lê-los e entendê-los, por fazerem parte de um rol de textos de leitura de certa forma obrigatória neste contexto acadêmico. Outros pontos que justificariam o uso desses textos são a facilidade de acesso e a diversidade de temas disponível. No entanto, percebeu-se que a extensão desses textos e sua concisão características podem apresentar tanto pontos positivos, quanto negativos na sua utilização em testes. Positivos, porque pode possibilitar maior rapidez de leitura e, desse modo, a utilização de mais de um texto, abordando assuntos diferentes, em um mesmo teste, caso haja interesse de se fazer um único teste para diversas áreas de estudo. E negativo, porque a concisão pode aumentar consideravelmente a complexidade desses textos e, com isso, dificultar a rapidez de leitura. Além disso, a utilização de vários textos sobre assuntos diferentes pode dificultar a resolução do teste, por exigir que o candidato possua e ative vocabulário técnico de áreas, às vezes, muito diferentes, em um espaço de tempo relativamente curto.

Dois fatores foram decisivos para que o uso desses textos fosse descartado nessa pesquisa. O primeiro ainda está relacionado com a extensão do texto. Os resumos de teses e dissertações são necessariamente concisos, como já foi dito; o que é um fator limitador do número e do tipo de questões possíveis de serem elaboradas. Em relação à limitação quanto ao número de questões, a autora entende que isso pode comprometer a confiabilidade do teste, pois não seria possível fazer um julgamento adequado da capacidade de leitura de um indivíduo por meio de duas ou três questões apenas e, embora

este seja um problema que pode ser solucionado com a utilização de mais de um texto da mesma área de estudos, há ainda a limitação quanto ao tipo de questão, que é bem mais significativa.

A limitação do tipo de questão é ocasionada pelo pouco detalhamento do assunto, característico desse gênero de texto, o que restringe as possibilidades de elaboração e faz com que grande parte das questões exija apenas uma recuperação de informações de fácil localização na estrutura superficial do texto. Essa tarefa pode de ser executada com a utilização de processo de leitura rápida e superficial, bastante comum entre leitores proficientes, e não pode ser considerado um problema em si mesmo. Entretanto, mesmo sendo parte fundamental do processo de compreensão de textos e, por isso mesmo, uma habilidade a ser avaliada, não é um tipo de leitura suficiente para avaliar a habilidade de compreensão de textos, no que concerne às estruturas mais profundas e complexas dos mesmos.

O segundo fator está relacionado com a qualidade dos textos pesquisados. Infelizmente, a qualidade desse tipo de texto não é a que se espera encontrar em trabalhos acadêmicos. No site da CAPES há uma quantidade imensa de textos, mas muitos deles são bastante difíceis de entender, devido a diferentes tipos de erro gramatical e má estruturação textual, além de falta de coesão e coerência[112]. Não raras vezes, até mesmo os textos em português (LM do escritor) são difíceis de entender, devido aos mesmos problemas citados acima.

Embora os pós-graduandos tenham que lidar com esses textos, de uma forma ou de outra, durante seus estudos, entendo não ser justo apresentar aos candidatos textos de qualidade duvidosa em um processo de seleção. Da mesma forma, seria contraproducente que os elaboradores de um teste tivessem muitas vezes que empregar mais tempo na edição ou reescrita desses textos, do que na elaboração do próprio teste.

Mesmo com estes problemas, elaborei quatro testes a partir de quatro resumos de teses[113], com a intenção de verificar na prática os problemas que esses textos poderiam apresentar se utilizados em uma avaliação. Os textos

[112]Ver exemplos de resumos no anexo 6.

[113] Ver apêndice 1.

escolhidos foram aqueles que apresentavam menos problemas de estrutura sintática e mais possibilidade de questionamento. Dos quatro testes, três foram realizados por um pequeno número de alunos do curso de Leitura para Compreensão de Textos do CELIN (Centro de Línguas e Culturalidade – UFPR), cada um em um momento diferente. Pude verificar que, de fato, os resultados não foram bons, conforme o previsto, principalmente devido à concisão e complexidade dos textos. Embora os alunos tenham sido capazes de responder as questões e que as respostas estivessem dentro de um padrão esperado de acertos e erros, o feedback oral dado pelos alunos e pela professora que aplicou o teste foi, de maneira geral, negativo, com comentários similares aos argumentos mencionados anteriormente.

O texto 1 é um exemplo de texto cujas perguntas possíveis demandam muito mais a capacidade de retirada de informações, do que capacidade de compreensão de texto. A média de acerto entre os alunos ficou em torno de 80%, o que indica que esse é um tipo de atividade que não exige muito dos candidatos e, portanto, acaba não discriminando quem é mais ou menos capaz.

O texto 3 é a respeito de assistência financeira para moradia. O problema nesse caso é que esse tipo de texto trata de uma questão regional, de recursos financeiros que pode não ter paralelo no sistema financeiro conhecido pelo candidato, dificultando a compreensão.

O texto 4 exigia que os alunos tivessem algum conhecimento de história, do vocabulário específico, ou que fossem capazes de usar seu conhecimento de mundo para fazer as inferências cabíveis. Algumas questões deste texto, por exemplo, obtiveram respostas com significados implícitos diferentes, que dificultam a correção e, muitas vezes, não permitem uma conclusão sobre quanto do texto foi realmente compreendido e quanto foi apenas literalmente traduzido, como os exemplos de respostas a seguir:

Pergunta 1: Que “instituições” eram essas?

Respostas:

a) eram instituições que davam assistência desemprego; criadas pelo governo canadense;

b) eram campos de assistência aos desempregados;
c) que iriam auxiliar o problema do desemprego;
d) grupos de assistência aos “desempregados” sob o comando do Departamento Nacional de Defesa do governo federal;
e) instituições de assistências no campo, conhecidas como instituições que oferecem subsídios para manter o homem no campo;
f) acampamentos de auxílio (albergue);
g) acampamentos de relevo do desemprego.

Pergunta 5: Qual era a opinião sobre essas “instituições” no oeste do país?

Respostas:

a) no oeste do Canadá, estas instituições chegaram a ser vistas como casa, “albergue” de inquietação (bagunça);
b) esses campos eram semelhantes a um antro de desocupados.
c) que estas instituições tem sido as camas quentes para a inquietação;
d) os grupos eram vistos em canteiros de agitação sem sossego, que não se ajustavam a este estilo de vida;
e) eram vistas como cama quente para o desassossego. Ou seja era vista como uma acomodação para estes homens;
f) que eram más-moradias, ou suspeitas.

As dificuldades encontradas no trabalho com os resumos mostraram que embora alunos de mestrado e doutorado tenham que saber lidar com esse gênero, ele não é uma boa alternativa para uso em avaliações de leitura a que nos propomos elaborar. A partir disso, busquei encontrar gêneros de textos mais confiáveis em termos de qualidade estrutural de língua e de articulação do discurso, que fossem de fácil acesso, oferecessem uma boa variedade de assuntos e que fosse relevante para as pesquisas e estudos desse meio acadêmico com o qual estamos trabalhando.

4.2.4.2 Textos Científicos

Na utilização de textos científicos, foram consideradas algumas variáveis, quais sejam: a extensão do texto, a fonte onde estes textos poderiam ser selecionados e a área de estudos à qual eles deveriam se referir preferencialmente.

No que diz respeito à extensão do texto foram consideradas três possibilidades: a primeira possibilidade foi a utilização de textos não muito longos, por questões práticas de tempo e de custo. O argumento que justificaria esta opção seria de que quanto maior o texto, mais tempo seria despendido na leitura e, conseqüentemente, mais longa deveria ser a prova; mais material e pessoal deveria ser empregado, elevando assim os custos do processo seletivo. Sob esse ponto de vista, então, haveria a necessidade de edição dos textos mais longos para deixá-los menores e, portanto, mais adequados segundo as limitações acima. Além disso, seria desejável que o texto mantivesse o máximo de fidelidade possível em relação ao texto integral. Assim sendo, a edição deveria ser feita com o mínimo necessário de alterações estruturais; eliminando o conteúdo que fosse considerado desnecessário no momento e mudando a estrutura apenas o suficiente para que a coesão e a coerência do texto fossem preservadas, tomando os devidos cuidados para que nenhum sentido ou informação do texto fosse corrompido.

A segunda possibilidade foi a utilização de apenas uma parte do texto. Ou seja, apenas a introdução, a metodologia, a discussão ou a conclusão, quando a parte em questão fosse abrangente o bastante para possibilitar a manutenção do enfoque do teste no uso de questões gerais básicas[114]. Isso estaria de acordo com a argumentação acima, com a vantagem de não precisar de edição alguma, possibilitando maior rapidez na elaboração do teste.

Essas duas possibilidades acima são utilizadas em diferentes contextos, incluindo testes de suficiência e podem gerar bons resultados. Entretanto, durante o processo de escolha de textos, surgiu um questionamento que levou a autora a rever os argumentos de escolha do texto

[114]Ver item 4.2.5 que trata da discussão sobre as questões propostas.

acima e influenciou a decisão final sobre a extensão dos textos a serem utilizados. A questão é: por que editar, ou usar partes de um texto, quando na prática os alunos de mestrado e doutorado têm que saber encontrar as informações de que precisam em textos originais, sem edição e de extensão variada? O contexto de pesquisa em que os alunos de mestrado e doutorado se inserem pressupõe a habilidade de seleção, compreensão e utilização de textos ou partes deles, que sejam necessárias à sua própria pesquisa. No estudo que estes alunos fazem, e em tanto outros, a leitura de um grande número de textos de extensão variada exige que o leitor saiba fazer a seleção de quais textos devem, de fato, ser lidos e saiba como e onde, dentro desses textos, eles podem encontrar as informações de que precisam, sem que para isso tenham que ler todo e cada texto na íntegra. Essa habilidade faz parte das habilidades de leitura de um leitor proficiente. Alderson afirma que

> um argumento comum em favor do uso de textos mais longos em, por exemplo, testes para propósitos acadêmicos, é que esta prática reflete mais de perto a situação onde estudantes têm que ler e estudar textos longos. Assim, mesmo se pesquisas ainda tenham que mostrar que certas habilidades podem somente ser avaliadas usando textos mais longos, o argumento de autenticidade mostra-se a favor do uso de textos mais longos, uma prática seguida pelo IELTS, por exemplo, em contraste com aquela do TOEFL, em que passagens mais curtas são usadas."[115] (2000, p. 109, tradução nossa).

Além do argumento que a própria questão coloca, existem ainda as vantagens de não haver necessidade de edição alguma e de haver uma grande variedade de fontes de textos científicos e de temas possíveis. Finalmente, ainda há as vantagens de que o leitor seja exposto a publicações acadêmicas e seja avaliado em relação à sua capacidade de trabalhar com textos, cuja extensão se assemelha aos quais ele, muito provavelmente, se defrontará no decorrer da sua pesquisa. Esse fato contribui sobremaneira, na opinião da

---

[115] A common argument in favour of the use of longer texts in, for example, testing for academic purposes, is that this practice reflects more closely the situation where students have to read and study long texts. Thus, even if research has yet to show that certain abilities can only be assessed using longer texts, the authenticity argument runs in favour of using longer texts, a practice followed by IELTS, for example, in contrast with that of TOEFL, where short passages are used.

autora, para que a avaliação tenha maior validade e mais confiabilidade. Dessa forma ficou decidida a utilização do texto integral, tal como é publicado, na elaboração do teste.

No que se relaciona à escolha das fontes de onde seriam retirados os textos, a autora considerou que jornais conceituados como *The New York Times* e *Guardian* e revistas científicas como *New Scientist* e *PNAS (Proceedings of the National Academy of Sciences)*, entre outros jornais e revistas online seriam fontes seguras, devido à idoneidade a elas reputada e à sua bem sucedida utilização nos testes de suficiência elaborados e aplicados pela UFPR.

A última variável foi a definição da área de estudos para a qual deveriam estar voltados os textos. A opção por concentrar os textos escolhidos para os testes em uma área de conhecimento apenas deveu-se, em especial, a dois fatores. O primeiro fator foi institucional. Os testes foram desenvolvidos com base nos padrões de testes aplicados na UFPR, que divide as provas de suficiência em três áreas de estudo, entre as quais os candidatos podem escolher para fazer o teste (em geral a escolha é feita de acordo com a área na qual ele pretende desenvolver seus estudos): Biológica, Tecnológica e Humana. O segundo fator foi de ordem prática: escolher apenas uma área entre as três acima tornou o trabalho de escolha de texto, de elaboração de questões e conseqüentemente de correção mais ágil e fácil, tanto para a elaboradora do teste quanto para os corretores, por delimitar mais o foco.

Em razão de ser possível encontrar uma quantidade maior de textos, cujos temas são acessíveis às três áreas de estudo, foi escolhida a área de Humanas[116] para a realização desta pesquisa. Dentro desta área foram selecionados vários textos de assuntos diversos, em uma primeira etapa. Em uma segunda etapa, foram escolhidos nove (09) textos, para os quais foram elaboradas algumas perguntas e respectivas respostas. Os nove textos e as respectivas questões e respostas foram analisados mais detalhadamente e foram selecionados, então, os três textos que foram utilizados no teste piloto[117].

a) Texto 1 - Prenatal thumb sucking is related to postnatal handedness

[116] Ver justificativa no início do item 4.1 acima.

[117] Ver os textos e os respectivos testes no apêndice 2.

- Fonte – site: Science Direct – Journal: Neuropsychologia, volume 43 – Issue 3 – pp.313-315 (2005)
- *http://www.sciencedirect.com/science/journal/00283932*
- Gênero[118] – Texto Formativo: Artigo de pesquisa

b) Texto 2 - Insect cells for human food

- Fonte – site: Science Direct – Journal: Biotechnology Advances, volume 26 – issue 5 – pp. 389-502 (September-October 2008)
- *http://www.sciencedirect.com/science/journal/07349750*
- Gênero – Texto Formativo: Artigo de pesquisa.

c) Texto 3 - In Diabetes, a Complex of Causes

- Fonte – Newspaper: New York Times - Published: October 16, 2007
- http://www.nytimes.com/
- Gênero – Texto para popularização da informação científica: Artigo de reportagem

Embora estes três textos tratem de assuntos relacionados à área biológica, o enfoque deles é dado em relação às questões sociais e antropológicas, o que justifica sua classificação como pertencentes à área de ciências humanas.

4.2.5 Elaboração das Questões

No desenvolvimento das questões a primeira decisão a ser tomada foi a língua que deveria ser usada: LE ou LM. Nuttall considera que

---

[118] Esta denominação de gênero, bem como as duas seguintes, segue a classificação apresentada por Goldman e Bisanz (2002) - in *Toward a Functional Analysis of Scientific Genres: Implications for Understanding and Learning Processes.*(ver referência completa na bibliografia)

> não é sempre possível expressar a questão que você quer perguntar em linguagem direta e clara, especialmente se o próprio texto é difícil. As pessoas algumas vezes sustentam que ler as questões é parte da tarefa de leitura, mas isto é só parcialmente válido. Esta é certamente *uma* tarefa de leitura, mas *a* tarefa de leitura é entender o próprio texto, e nós poderíamos argumentar que qualquer coisa que desvia disso não ajuda.[119] (2000, p. 187)

Seguindo esta linha de argumentação, a língua escolhida para a formulação das questões do teste foi a LM, i.e, a língua portuguesa, na medida em que a idéia não é saber se os leitores são capazes de entender as questões, mas de respondê-las. As questões, por isso, não podem se constituir em um desafio de interpretação a mais para o leitor; elas devem ser tão claras e compreensíveis quanto possível, permitindo que o leitor utilize seu tempo e concentração na leitura e entendimento do texto.

Nery (2003, p. 49) constrói uma matriz de questões que considera os tipos de texto, no que concerne à forma de apresentação: *informativo* e *argumentativo,* e o grau de compreensão que se exige do leitor. Nesta matriz ela propõe seis tipos básicos de questão, já apresentados no item que trata do leitor ideal/leitor real e que serão reapresentados a seguir:

a) *reconstituição da informação* – que exigem que o leitor identifique e extraia as informações solicitadas, tais como aparecem no texto;
b) *ordenação e relevância* – que exigem que o leitor reconstitua a ordenação das informações que se articulam no texto, considerando, inclusive o grau de relevância de cada uma delas;
c) *estabelecimento de relações* – que exigem do leitor a apreensão de relações existentes entre elementos do texto, partes do texto e textos entre si;
d) *reconhecimento do quadro enunciativo* – que exigem que o leitor compreenda as estratégias discursivas em função dos componentes da cena enunciativa;

---

[119]"It is not always possible to express the question you want to ask in straightforward language, especially if the text is itself difficult. People sometimes maintain that reading the questions is part of the reading task, but this is only partly valid. It is certainly *a* reading task, but *the* reading task is making sense of the text itself, and we could argue that anything that distracts from this is unhelpful."

e) *apreensão de julgamento de valor* – que exigem que o leitor reconheça e compreenda os julgamentos de valor veiculados no texto;

f) *reconstrução da argumentação* – que exigem que o leitor seja capaz de reconstruir a linha argumentativa que através da qual a informação é articulada.

Cada um desses seis tipos se subdivide em três outras categorias possíveis, quais sejam: *orientadas* ou *não orientadas,* quando possuem ou não algum tipo de orientação de leitura; *pontuais* ou *globais,* quando se referem a informações mais localizadas dentro do texto ou que são apresentadas de forma mais global no texto, e *lineares* ou *não lineares,* quando as informações podem ser localizadas em um ponto específico e único do texto ou em vários pontos ao longo dele todo. Segundo ela "o cruzamento de cada um desses critérios permite a construção 'teórica' de diferentes tipos de questões[120]. O grau de complexidade das questões, segundo ela, está distribuído entre dois pólos que vão do mais informativo para o mais argumentativo, em que as questões mais simples seriam de *reconstituição da informação pontual linear* e as mais complexas as de *reconstrução da argumentação não orientada global*, podendo variar ao longo de um continuum a depender das combinações entre os critérios possíveis.

Entre os tipos de questão classificados por Nery, os itens que são mais significativos, dos seis acima, para esta pesquisa são a reconstituição da informação, apreensão de julgamento de valor e reconstrução da argumentação, incluindo as subdivisões das outras três categorias propostas. Isso não significa que todos esses tipos de questão são utilizados em cada teste. A elaboração de um ou de outro tipo de questão depende das possibilidades e limitações impostas pelo próprio texto, além dos propósitos traçados de utilização de questões que enfoquem o entendimento de informações recorrentes em trabalhos de pesquisa científica.

Como mencionado anteriormente, a proposta desse trabalho é tentar manter nos testes de suficiência a utilização de questões básicas. Neste

[120] Para maiores detalhes sobre a classificação de Nery indicamos a leitura do livro Questões Sobre Questões de Leitura, cuja referência completa se encontra na bibliografia.

trabalho, entende-se por questões básicas, aquelas questões as quais são comumente feitas pelo próprio leitor; que muitas vezes levam o leitor a ler determinados textos; que constituem o esqueleto de qualquer pesquisa, e que são fundamentais na compreensão da leitura de textos científicos e acadêmicos em geral: o que, quem, qual, onde, por que e como. São questionamentos que tornam possível a compreensão da idéia geral e das principais informações do texto, que podem ser distribuídos em quatro tópicos, que constituíram o foco das questões, quais sejam: o *objetivo*, a *metodologia*, os *resultados* e a *conclusão* da pesquisa ou do trabalho apresentado no texto.

Considerando o foco desejado, foram propostas as seguintes questões gerais, que serviram de base para a escolha dos textos:

1- Qual é o objetivo da pesquisa / experimento?
2- Qual foi o procedimento / método utilizado na pesquisa / experimento?
3- A que resultados os pesquisadores chegaram?
   Ou: Quais foram os resultados obtidos?
4- Foram traçados novos objetivos a partir dos resultados? Se foram, quais?
5- Qual a conclusão do(s) autor(es)?

A utilização desse tipo de questão como base para escolha de textos, não visa limitar a escolha, mas direcioná-la, no sentido de buscar textos em que esses tópicos estejam presentes de forma mais clara e explícita, pois a autora defende a proposta de que o teste não deve conter “armadilhas” ou “pegadinhas” para o leitor, ele deve ser uma forma justa, tanto para este quanto para a instituição, de garantir que o indivíduo possui os conhecimentos que serão necessários durante o seu estudo. Sendo assim, o texto cuja organização estrutural é mais complexa, e que pode causar muitas dúvidas ou gerar ambigüidade não deve ser utilizado, evitando possíveis “armadilhas” discursivas, que prejudiquem os leitores. Os textos selecionados para os testes desta pesquisa estão conforme a proposta acima e, como se verificará adiante, o resultado dessas escolhas foi positivo, pois facilitou o trabalho de elaboração dos testes e, pela observação dos resultados, conferiu maior confiabilidade, pela discriminação que possibilitou.

Dizer isso, no entanto, não significa dizer que as avaliações ficaram isentas de quaisquer problemas. Como toda avaliação, esta também os teve, no que diz respeito à elaboração das questões, como poderá se observar na análise dos testes.

Após a definição dos três textos utilizados e a partir das questões acima, foram reelaboradas, então, as questões de cada texto, respeitando as características de cada um e procurando manter o foco nas questões mais relevantes dos textos. Isso resultou em um número de questões diferente para cada texto, em função da abordagem específica de cada um deles, sendo:

- Texto 1 (piloto A) – 07 questões
- Texto 2 (piloto B) – 08 questões
- Texto 3 (piloto C) – 05 questões

Seguindo a proposta de elaboração das questões apresentada, pode-se perceber que o número de questões do texto 3 ficou bastante reduzido. Isso poderia representar, como confirmado posteriormente, um problema na realização e também na correção, colocando em risco a confiabilidade deste teste. Entretanto, em lugar de eliminar o texto, optou-se por mantê-lo, a fim de que se pudesse observar a reação dos leitores àquelas questões globais e abrangentes em termos de conteúdo, e elaborar um outro conjunto de questões, que, embora mantivessem o foco já determinado, fossem mais específicas e incluíssem mais detalhes, o que abriu a possibilidade de obter um maior número de questões, bem menos abrangentes, por serem mais locais. Mantendo, então, os mesmos três textos, o número de testes aumentou para quatro:

— Texto 3 (piloto D) - 10 questões

Em resumo, a organização dos testes em termos numéricos ficou da seguinte maneira:

— Número de testes-piloto: quatro (04)

— Número de questões por teste: piloto A (07) / B (08) / C (05) / D (10)

Da mesma forma que se tentou manter a uniformidade das questões nos quatro testes, também se tentou manter a distribuição de mesmo grau de dificuldade nas questões dos diferentes testes, para que eles não

apresentassem graus de dificuldade global muito desiguais. No entanto o teste C, foi a exceção (prevista desde a elaboração das questões) e, além disso, essa distribuição de grau variou em função das limitações impostas pelo assunto, da estrutura dos textos e da forma como a própria questão foi elaborada.

Os testes contêm questões de diferentes graus de dificuldade, porque as questões não podem ser só difíceis ou só fáceis; um teste válido deve, entre outras coisas, conter questões de nível de dificuldade variada: difíceis, médias e fáceis, ou não será possível se fazer a discriminação entre os diferentes níveis de habilidade dos leitores. Segundo Bachman (1990, p. 207) "um item difícil vai fornecer muito pouca informação em níveis de habilidade baixos, uma vez que virtualmente todos os indivíduos neste nível vão errar o item. Igualmente, um item fácil que indivíduos de habilidade elevada respondem corretamente de maneira uniforme não vão fornecer muita informação neste nível de habilidade." No entanto, muitas vezes o próprio texto impõe certos limites à formulação de questões, fazendo com que na prática as questões de alguns testes fiquem mais difíceis ou mais fáceis que de outros[121].

As questões foram organizadas e apresentadas de acordo com a ordem em que os assuntos aparecem no texto. A elaboração das questões respeita o objetivo de avaliar a compreensão e pontos básicos apresentados nesses gêneros de texto. Por exemplo: o que, por que, quando, como, etc, conforme explicação mais detalhada acima.

Finalmente, também foi levado em consideração na elaboração das questões o fato de que, especialmente quando são utilizadas questões discursivas no teste como neste caso, além da variabilidade em função da capacidade de compreensão e habilidade de expressão do candidato, há ainda outra fonte de variação muito importante: o julgamento subjetivo, por parte do corretor, que inevitavelmente acontece no ato da correção.

[121] Ver discussão sobre a elaboração dos testes-piloto 1C e 1D no capítulo V.

4.3 CORREÇÃO DOS TESTES-PILOTO

A correção dos testes-piloto foi realizada, propositalmente, sem estabelecimento de critérios para a utilização em comum entre os corretores[122], para que fosse possível analisar os problemas de correção a partir de dados gerados nesta pesquisa e observar os resultados fornecidos pelo aplicativo.

- Número total de corretores: dois (02)
- Número de corretores por teste: dois (02)
- Número de total de testes aplicados: noventa e quatro (94)
- Número de testes aplicados por teste: piloto A (18) / B (26) / C (24) / D (26)

Como o número de questões de cada teste era diferente, cada questão foi pontuada de forma diferente:

Cada questão foi pontuada de 0 a 10, e depois foi feita a equivalência da nota geral do teste para que todos os testes tivessem pontuação equivalente. Por exemplo: um candidato que tivesse realizado o piloto A, cujas notas das questões de 1 a 7 fossem, respectivamente 6; 4; 8; 7; 7; 6 e 5, atingiria 43 pontos dos 70 possíveis. Fazendo a pontuação total do teste equivaler a 100 pontos, com uma regra de três simples, se obteria a pontuação equivalente de 61 pontos, para o candidato em questão.

Para apresentar os resultados obtidos na realização e correção dos testes-piloto foi feito um levantamento dos problemas que surgiram na elaboração das questões, nas respostas dadas e na correção feita por dois avaliadores. Com base nos problemas encontrados algumas soluções foram propostas, para que, posteriormente, um dos testes fosse refeito e reaplicado.

---

[122] Isso não significa que nenhum critério tenha sido estabelecido. Cada corretor estabeleceu e usou seus próprios critérios, com base em sua experiência docente no ensino de leitura em ESP.

4.3.1 – Problemas Identificados na Elaboração das Questões

As questões dos quatro testes foram elaboradas com base nas questões gerais básicas, apresentadas no final do capítulo VI. Durante o processo levamos em consideração as necessidades de adaptação das questões, de acordo com as características e as informações presentes em cada texto. Como foi dito anteriormente, as questões gerais serviram apenas de base, visto que não seria possível aplicar literalmente as referidas questões, em todo e qualquer texto, mesmo que de um mesmo gênero, devido ao assunto, à forma e profundidade dados a cada texto.

Foram identificados nos teste os seguintes problemas com as questões:[123]

No **piloto 1A** a questão 2 (*Que metodologia os pesquisadores utilizaram?*) parece não ter ficado clara. Alguns respondentes tiveram dificuldade em sintetizar o item 2 do texto, que trata do método.

Resp1:. *Foram observados 80 fetos (e seus pais) durante o período da gestação. Após os pais foram consultados para ver se queriam fazer parte da continuação do estudo. Em que deveriam responder a um questionário sobre atividades das crianças e com que mão ou pé realizavam estas atividades.*

75 concordaram em participar. Destas 75, 60 sugavam o dedo direito na gestação e 15 o esquerdo.

Resp2:. *Uma metodologia estatística.*

Resp3:. *Observaram 75 crianças. Destas 60 chupavam seus dedos da mão direita e 15 seus esquerdos.*

Uma possível causa da confusão talvez tenha sido gerada pela pergunta 3, que fala de uma “segunda fase”. Este parece ser o caso observado na resposta 2 acima. Outra possibilidade é que o conceito de “metodologia” não

[123] Para facilitar o acompanhamento dos problemas levantados, vamos estabelecer a ordem dos comentários de acordo com a ordem dos pilotos (de A , C e D), sendo que as questões que não apresentaram problemas, não serão mencionadas. O teste piloto 1B será devidamente comentado no capítulo em que propomos a sua reformulação.

esteja claro para esses respondentes, porque se a pergunta 2 se referisse à pesquisa realizada no período gestacional, seria totalmente incoerente, visto que o texto apenas menciona a pesquisa anterior para contextualizar a pesquisa em questão. O texto não apresenta quaisquer detalhes metodológicos do primeiro estudo. Observamos, portanto, que a frase "segunda fase", que aparece na questão 3 poderia causar confusão, mais provavelmente, naqueles leitores desatentos, ou que apresentassem o desconhecimento conceitual citado acima.

Outra causa ainda, é a generalização feita pela questão. Como a metodologia está subdividida em três partes (participantes, procedimento e análise e resultados), isso talvez possa ter gerado uma dúvida sobre o quão abrangente deveria ser a resposta.

A solução para este problema seria reformular as questões 2 e 3, para deixá-las mais claras e objetivas, da seguinte forma:

2) Que metodologia os pesquisadores utilizaram na pesquisa, no que diz respeito aos participantes?

3) Qual o objetivo das dez perguntas feitas aos sujeitos?

4) Por que razão uma das questões foi desconsiderada?

Dessa forma, a questão 2 seria mais pontual e a questão 3, que trata dos procedimentos metodológicos, não traria o enunciado inicial.

A questão 6 foi mal formulada. Não havia necessidade de dividi-la daquela maneira. Uma possível solução seria a seguinte reformulação: *A pesquisa sugere que nem todos os itens mostraram-se adequados na indicação de "handedness". Dê um exemplo e explique por que.*

No teste **piloto 1C** foi utilizado um artigo de divulgação científica do jornal New York Times, que mostra que o diabetes deve ser atribuído a um complexo de causas e não a uma causa única. No entanto, por ser um artigo de divulgação em jornal, a distribuição dos tópicos enfocados nas questões propostas não está ordenada seqüencialmente como geralmente acontece em textos de divulgação científica em revistas especializadas e publicações acadêmicas. Este fato, como nós imaginamos durante o processo de elaboração das questões, gerou o que podemos considerar um problema no teste todo, que é a exigência de uma grande capacidade de organização e

síntese de informações por parte do respondente. Em especial na questão 2, que pergunta: *Quais as hipóteses levantadas nas diferentes pesquisas mencionadas no texto?*. A resposta para esta pergunta está bastante distribuída, porque várias hipóteses são apresentadas ao longo de todo o texto, exigindo que o leitor faça uma compreensão geral do texto e que seja capaz de identificar especificamente cada hipótese levantada, sem confundir as hipóteses com os resultados já verificados das pesquisas. O leitor deveria, então, ser capaz de sintetizar de forma coerente cada hipótese para formular a resposta.

O índice de confiabilidade do teste C foi o mais baixo de todos. Entendemos que isso pode ter acontecido em função das próprias questões, em função das respostas, ou em função da correção. Se analisarmos a explicação acima, do porque as questões podem ter sido o problema, veremos que, na verdade, um leitor proficiente deveria ser capaz de respondê-las sem maiores problemas. Isto significa dizer que este é o nível de proficiência de leitura de texto que podemos considerar ideal. Um leitor capaz de organizar e sintetizar informações dispersas em um texto como o utilizado nesse piloto, provavelmente será capaz de compreender outros textos em que as informações estejam dispostas mais linearmente. O problema desse teste se deve, na verdade, ao contexto educacional do momento. O fato é que não podemos contar com leitores com tal proficiência, nem mesmo em LM. Podemos observar isso com mais precisão e clareza em sala de aula, quando os alunos têm chance de responder questões como essas na escrita e depois oralmente. Percebemos que embora o aluno tenha compreendido e identificado as informações necessárias, ele não consegue, muitas vezes, organizar, conectar e sintetizar essas informações de maneira coerente, principalmente na escrita. Como a redação da resposta é feita em LM, podemos perceber melhor a relação que existe entre a proficiência na leitura e a capacidade de redação. Embora não tenhamos dados concretos para fazer tal afirmação, e nem seja esse o nosso objetivo neste trabalho, nossa experiência em sala de aula nos mostra que aqueles alunos que são leitores proficientes em LM, mesmo que não entendam perfeitamente um texto em LE, são capazes de redigir respostas mais consistentes, do que aqueles menos proficientes, que

obtiveram o mesmo grau de entendimento do texto. Esta, portanto, pode ter sido a causa dos problemas encontrados nesse piloto. E, em razão de estarem envolvidos tantos fatores extras, esse teste foi desconsiderado para a continuação do processo de elaboração do teste de suficiência, pois não haveria tempo suficiente para que fossem feitos todos os estudos necessários de todas as variáveis envolvidas.

No piloto 1D o texto utilizado foi o mesmo do teste anterior. O objetivo era tentar elaborar questões que estivessem baseadas nos mesmos tópicos, mas fossem menos abrangentes que as do piloto 1C, para que pudéssemos comparar e avaliar os prós e contras da utilização desse gênero, com os dois tipos de questões (mais gerais X mais específicas). Acabamos de analisar o que ocorreu com as questões mais gerais, para as quais não propusemos soluções, pois uma das soluções possíveis seria utilizar outro tipo de questão como fizemos neste teste que vamos analisar.

Os principais problemas encontrados foram com relação à redação das questões. Na questão 4, *Que idéia de um antigo pesquisador serviu de motivação para a pesquisa?,* por ser bastante específica, parece ter causado alguma confusão quanto à qual pesquisa a pergunta se referia. O único pesquisador antigo citado no texto é Claude Bernard, mas talvez ficasse mais claro se a pergunta explicitasse a que parte da pergunta ela se referia.

A questão 6, *Que papel o esqueleto desempenha na questão do diabetes?* também causou problemas, porque vária pessoas não entenderam o emprego do termo “esqueleto” como sinônimo de “ossos”. Além disso, essa questão poderia ter sido complementada, nela mesma ou com outra pergunta, pedindo que fosse relatada qual a conseqüência da deficiência da substância produzida pelos ossos.

A questão 7, *Qual é o papel desempenhado pelo cérebro?*, é de certa forma redundante, pois as informações podem ser incluídas na questão 2, *Quais as novas descobertas sobre diabetes?*. A questão 9, *Que outro órgão, mencionado no final do texto, também está envolvido na regulagem do açúcar? Como?*, embora não tenha gerado problemas relevantes, evidenciou o problema recorrente de desconhecimento dos respondentes de que os intestinos constituem um órgão. A questão 10, *Essas descobertas e o*

desenvolvimento das novas drogas são conclusivas? Por que?, gerou certa confusão. Várias pessoas responderam afirmativamente, quando, na verdade, foram levantadas várias hipóteses e a autora afirma a necessidade de mais estudos textualmente. Talvez ela ficasse mais clara se a pergunta deixasse mais explicito que se referia às informações da conclusão do texto.

A autora admite a possibilidade de que as soluções propostas para os problemas encontrados nas questões desses três testes talvez não sejam as melhores. Essas são apenas soluções dadas rapidamente a partir dos problemas levantados, mas que não foram colocadas em teste ou analisadas em profundidade e que, por isso, não podem ser vistas como soluções definitivas, mas apenas como sugestões de solução.

4.3.2 Problemas Observados nas Respostas

Na análise, comprovou-se a interferência da redação das respostas na correção, conforme previsto anteriormente. Os problemas de redação em LM são muitos e alguns bastante relevantes para nossa análise. Há inúmeros problemas de ortografia, pontuação, acentuação, concordância, sintaxe e, o que é mais problemático, problemas de coesão e coerência. Como exemplo de alguns desses problemas, são transcritas algumas perguntas e respostas abaixo.

Teste 1A

4) Quais foram os resultados obtidos?

Ma: *O estudo indicou que o comportamento motor **lateralizado**, neste caso a **suquição** do polegar está muitíssimo relacionada com a lateralidade quando na vida adulta (destro ou canhoto), talvez mais fortemente para fetos destros do que fetos destros. As evidências se mostraram mais contundentes para encontrar uma relação entre a **suquição** do polegar da mão direita durante o período fetal e a manifestação de **destralidade** do que para os fetos que sugavam o polegar esquerdo.*

Ms: *As crianças que responderam com a mão direita tiveram score 4 ou poucos responderam e as crianças que responderam com a mão esquerda tiveram score de 5 ou mais.*

Os resultados indicaram o polegar direito no pré natal é subseqüente mão direita em 100%. Os associados ao polegar esquerdo é subseqüente a mão esquerda em 60%.

5) Na pesquisa, o sexo dos bebês também foi considerado. O que os pesquisadores observaram?

L: *Que o numero de* ***individuos*** *que pontuou de 9 para 0 quando usou a mão esquerda foi de acordo com o sexo. Os homens apresentaram mais uso da mão esquerda*

Teste 1C

2) Quais as hipóteses levantadas nas diferentes pesquisas mencionadas no texto?

SC: - *hormônios estão presentes em alguns locais do esqueleto, que irão auxiliar o corpo a controlar o açúcar.*

- Elevadas quantidades de açúcar no sangue são conseqüências da diabetes, sendo que varia entre as pessoas as razões desta taxas irregulares

- Relação de que a gordura regula os ossos e os ossos regulam a gordura (trabalhos com ratos demonstrou que uma substância produzida no osso é estimulada a atuar pelas células de gordura no pâncreas e este funcionamento auxilia em uma melhor secreção da insulina para o rato como também ajudar na circulação da glicose pelo sangue, interior das células, músculos). Insulina é um importante regulador de lipídios.

- Testes com aumento de osteocalcin, levaram a uma elevada produção de insulina e a quantidade de açúcar no sangue baixou

- Deficiência na regulação da glicose aciona o sistema imune, com o auxílio dos macrófagos, podem causar inflamação.

Algumas vezes as respostas mostraram-se bastante incoerentes ou ambíguas e, embora contivessem vocabulário e idéias do texto, não foi

possível ter certeza se o leitor havia de fato entendido ou não o texto em inglês, como acontece nos exemplos abaixo:

Teste 1A

2) Que metodologia os pesquisadores utilizaram?

NC: *Oitenta crianças foram observadas previamente quando eram fetos. Essas crianças foram* ***contactadas****, e convidadas a participar da seleção, através de questionários para participar da pesquisa. 75 entraram em acordo na data reportada aqui, outras não, duas não podiam se deslocar de casa e três não fizeram os exames. Quatro delas tinham o lado direito e 1 o lado esquerdo. Das 75 crianças que participaram 60* ***(?-está faltando o verbo)*** *o polegar direito e 15 o polegar esquerdo.*

MHP: - *Da relação hormonal, leptin, um hormônio produzido por gordura, um importante regulador do metabolismo ósseo, trabalhando com ratos, o hormônio osteocalcin, produzido nos ossos atua em células gordurosas bem como o pâncreas, ajudando a secretar insulina e regular a glicose.*

- O aumento deste hormônio ósseo pode ser a causa da diabete tipo 2, onde a insulina torna-se resistente, e a glicose no sangue no mesmo nível dela, produção de insulina no sangue declina. (relação

-Relação dos macrófagos (sistema imunológico) com a obesidade.

- Relação do cérebro para o controle da glicose.

GMF: - *Aplicação de questionários, onde das 80 crianças previamente observadas, 75 concordaram em participar do estudo, onde 60 usavam a mão direita e 15 a esquerda. As crianças que participaram desta pesquisa tinham de 10 a 12 anos.*

7) Esses resultados obtidos foram considerados conclusivos? Por que?

Ms: *Não. Foi previsto uma sugestão para os resultados padrão de comportamento e uma continuação da pesquisa em sexos femininos e*

masculinos. Ainda o teste sugere a continuidade entre as crianças sem mãos para 1º ano de vida e anos depois.

Teste 1C

1) Quais os objetivos da pesquisa apresentados no texto? (Explícitos e/ou implícitos)

MHP: *Diante da preocupação, da diabete ser uma doença apontada como o 5º assassino dos americanos, a pesquisa apresenta o que é a diabete, mas possíveis causas, e questionamento de como é regulada, de como o açúcar está mais ou menos presente no sangue. Apresenta novas pesquisas com a importância não somente do fígado, pâncreas, músculos e gordura. Aponta para o cérebro, intestino, sistema imunológico, hormônios dos ossos como descobertas p/ o controle da glicose e consequentemente ao tratamento da diabete, apresenta qual poderia ser a causa da insulina resistente.*

Algumas respostas denotam a falta de conhecimento do leitor do quê está sendo perguntado e dos conceitos envolvidos na questão. Por exemplo, na resposta da questão 2 do piloto 1A, transcrita a seguir, o leitor demonstra não ter conhecimento exato do que seja uma metodologia, formulando uma resposta, além disso, com informações incorretas, de acordo com o texto.

2) Que metodologia os pesquisadores utilizaram?

Resp1:. *A metodologia utilizada pelos pesquisadores teve sua origem de uma pesquisa realizada com 75 crianças, onde tiveram um questionario de 10 questões mais a estipulação de algumas tarefas, no qual receberiam notas de zero a nove. Todas as crianças tinham média de idade entre 10 e 12 anos.*

Resp2:. *Foi utilizado a metodologia da lateralização hemisférica, onde é estudado os lados do cérebro (direito esquerdo).*

Os problemas relacionados com as respostas acima não são o objeto deste estudo e, portanto, não será possível sugerir soluções para cada um deles, nesse momento. Entretanto, esta autora não poderia deixar de

apresentá-los, porque eles representam um fator de influência muito relevante no resultado dos testes.

Além dos problemas de correção, em todos os pilotos houve problemas de correção. Esses problemas são o reflexo daqueles mencionados no capítulo 2 item 2.12, que trata da correção de testes e dizem respeito especialmente à falta de estabelecimento de critérios de correção em comum, para os dois corretores. Entretanto, como os mesmos problemas se repetem em todos os testes, só foi feita a análise dos problemas de correção na análise dos resultados das questões do teste 1B, que é apresentada mais adiante.

4.4 ESCOLHA, REELABORAÇÃO E APLICAÇÃO DO TESTE

O resultado geral dos testes, de acordo com a análise feita com o Modelo Rasch, foi bastante positivo. A avaliação do modelo estabelece uma classificação do desempenho geral dos testes em 5 níveis, a saber: MUITO FRACO; FRACO; RAZOÁVEL; BOM e EXCELENTE. Os testes A e D foram classificados pelo modelo como “Excelente”. Conforme demonstrado acima na elaboração das questões, esses testes ainda apresentam problemas, mas como são em menor número em relação aos outros dois, provavelmente não gerarão resultados diferentes o suficiente para uma comparação e análise substancial. Os testes B e C obtiveram conceito de classificação “Bom”, apresentando um maior número de problemas na elaboração das questões. O objetivo da escolha de um teste apenas foi possibilitar a solução dos problemas encontrados nas questões e verificar a existência de outros problemas que podem, freqüentemente, ser encontrados em testes. Deste modo, foi possível realizar as necessárias alterações e, a partir das modificações realizadas, verificar se as modificações gerariam, ou não, maiores diferenças nos resultados, possibilitando uma verificação mais criteriosa da real utilidade de um aplicativo como o Modelo Rasch e da construção de uma escala de avaliação que possa servir de instrumento de equalização na análise de testes de suficiência.

Piloto 1A – Classificação: Excelente – Índice de confiabilidade: 0,87258

Piloto 1B – Classificação: Bom – Índice de confiabilidade: 0,79341

Piloto 1C – Classificação: Bom – Índice de confiabilidade: 0,76041

Piloto 1D – Classificação: Excelente – Índice de confiabilidade: 0,87000

Essa classificação do teste encontra-se no quadro que apresenta o resumo das estatísticas geradas pelo aplicativo e elas dependem dos dados fornecidos. No caso da classificação "Bom", do piloto 1B acima[124], o "poder de adequação do teste" como é denominado, (literalmente = *Power of test-of-fit*) depende do "índice de separação do indivíduo" (*person separation index*), que é um dos índices de confiabilidade disponíveis no modelo. Este índice varia de 0 a 1. Considera-se que quanto mais baixo é o índice de separação do indivíduo, i.e., quanto mais próximo de 0, maior a tendência de que os indivíduos estejam localizados próximos uns dos outros não sendo possível distinguir se os indivíduos com localização elevada tendem a obter escores mais altos que os indivíduos com baixa localização, ou não. Isso significa que o poder do teste é baixo, porque ele não permite que se estime a probabilidade de acerto em função da habilidade dos indivíduos com segurança. A partir disso, o poder de adequação do teste piloto 1B acima é considerado Bom, devido ao índice de separação do indivíduo estar mais próximo de 1 do que de 0 (0,79341).

4.4.1 Escolha do Teste e Re-elaboração do Teste

Entre os dois testes, B e C, foi feita a opção de se trabalhar com o teste B mais detalhadamente para aplicação como teste final, devido aos resultados

[124] Ver tabela "The Summary Statistics" deste piloto no apêndice 3.

obtidos na análise feita pelo aplicativo (Modelo Rasch), e devido ao fato de que as questões deste teste representavam mais fielmente o tipo de questão proposta inicialmente.

O objetivo da elaboração das questões, bem como do tipo de correção efetuada (sem estabelecimento de critérios e treinamento de corretores) e da análise dos resultados, não era a obtenção de um teste que pudesse ser considerado ideal, ou perfeito, considerando os conceitos e características apresentados ao longo desta dissertação. Ao contrário, procurou-se elaborar e corrigir os testes de acordo como, em geral, os testes são elaborados e corrigidos na grande maioria das instituições e situações atualmente, ou seja, com a intenção de elaborar um bom teste, mas sem a utilização de critérios muito bem definidos, porém, usando muito mais a intuição e a experiência docente, ao invés de basear-se em alguma teoria de avaliação e correção de testes existente.

A intenção foi, justamente, possibilitar o levantamento de problemas comuns às avaliações, para que pudessem ser identificados e analisados, na busca de soluções que pudessem minimizar esses problemas na elaboração e aplicação de futuros testes de seleção. Sendo assim, com base nos resultados obtidos, o objetivo foi reelaborar as questões que apresentaram maiores problemas, quer na elaboração das respostas, para os candidatos, quer na correção, para os corretores.

Na re-elaboração do teste, buscou-se identificar quais questões apresentavam problemas, para que fossem melhoradas ou, caso isso fosse necessário, substituídas por outras totalmente diferentes. Foram mantidas as questões: 1; 2; 4; 5 e 6, sendo que nessa última foi redefinida a formatação com a inclusão de um espaço específico indicado para a relação de vantagens e desvantagens, minimizando assim, problemas de correção devidos à desorganização espacial das respostas. As questões 3; 7 e 8 foram modificadas da forma apresentada abaixo, sendo que em negrito estão as questões já reformuladas e aplicadas no teste final:

3) De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, preencha a tabela abaixo.

Proteína	Sucesso ou Fracasso	Razão

TABELA 5 –TABELA USADA NO TESTE

FONTE: AUTOR (2008)

Razão: O formato desta questão demonstrou ser excessivamente facilitador, não possibilitando a discriminação desejada entre os candidatos mais fortes e mais fracos.

3) De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, escreva abaixo:

a) a denominação das 3(três) proteínas que foram apresentadas ao mercado consumidor;

b) a reação do mercado consumidor (se for apresentada) e a razão do sucesso ou fracasso de cada uma delas.

Expectativa: Que a junção das duas colunas da direita e a necessidade de identificação e argumentação da razão do fracasso ou sucesso de cada uma proporcionasse uma discriminação mais observável e relevante para o processo.

7) Os autores do artigo chegaram a que conclusão?

Razão: Assim como na questão 1, os candidatos em geral não conseguiram distinguir o que era conclusão de pesquisa do que era conclusão do artigo. Para que não houvesse duas questões gerando este mesmo tipo de problema, optamos por modificar esta, em que, aparentemente, o problema mostrou-se mais evidente.

7) Diante da dificuldade que os consumidores demonstram em incluir os insetos em seus cardápios, os pesquisadores pensam em uma outra possibilidade. Diga qual a possibilidade e explique por que eles a consideram viável.

Expectativa: Evitar o problema acima descrito. Esta questão foi trocada e não apenas modificada.

8) Qual é o papel do consumidor nessa história?

Razão: Esta questão ficou ambígua, possibilitando mais de uma resposta, que poderiam ser consideradas corretas.

8) Que argumentos podem justificar mais estudos em relação ao ponto de vista do consumidor?

Expectativa: Que a reformulação da questão eliminasse a ambigüidade.

Antes da aplicação do teste final, foi feito pela autora um julgamento da dificuldade das questões reelaboradas, para o qual foram definidos 3 graus de dificuldade: difícil, médio e fácil. O objetivo dessa classificação era um posterior confronto entre esta e a classificação com base nas informações geradas pelo aplicativo. O julgamento teve como base o tipo de informação requerida, considerando a classificação teórica apresentada por Nery e nas dificuldades identificadas nas respostas durante a correção do piloto. No entanto, como a análise da dificuldade das questões pretende ser mais embasada nas informações geradas pelo modelo Rasch do que por modelos teóricos de classificação, para cada questão foi proposto um dos três graus de dificuldade, com breve justificativa da escolha.

Q1- Qual é o objetivo da pesquisa?

DIFÍCIL – Exige o uso de um conhecimento prévio da existência de uma diferença entre o objetivo da pesquisa do objetivo do artigo e o posterior reconhecimento do objetivo solicitado.

Q2 – O que motivou a pesquisa? Explique.

FÁCIL – Exige a reconstituição de uma informação que se encontra bastante explícita no início do texto.

Q3- De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, escreva abaixo:

1. a denominação das 3(três) proteínas que foram apresentadas ao mercado consumidor;
2. a reação do mercado consumidor (se for apresentada) e a razão do sucesso ou fracasso de cada uma delas.

1:FÁCIL – Exige apenas a identificação da informação, que é dada explícita e linearmente no texto.

2:MÉDIA – Cada informação encontra-se imediatamente após a informação do item 1 acima, portanto, fácil de ser recuperada. No entanto as informações são apresentadas de forma menos explícita.

Q4- Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento.

MÉDIA – Exige o reconhecimento da estrutura de contraste e recuperação dela na elaboração da resposta, o que em geral dificulta a questão; porém a informação é orientada, o que minimiza esse grau de dificuldade.

Q5- Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína?

FÁCIL – Exige a reconstituição da informação, que é fornecida no texto de forma pontual e linear.

Q6- Quais as vantagens e desvantagens da utilização da nova proteína como alimento?

MÉDIA – Embora seja uma questão que também exija o conhecimento da estrutura de contraste, como a Q4, esta estrutura é apresentada de maneira bastante explícita no texto. O que a torna mais difícil é o fato de ser uma questão trabalhosa, por exigir a recuperação de várias informações, cuja linearidade se confunde um pouco na apresentação de dados numéricos.

Q7- Diante da dificuldade que os consumidores demonstram em incluir os insetos em seus cardápios, os pesquisadores pensam em outra

possibilidade. Diga qual a possibilidade e explique por que eles a consideram viável.

DIFÍCIL – Exige a recuperação da informação pedida e a reconstrução da argumentação apresentada.

Q8- Que argumentos podem justificar mais estudos em relação ao ponto de vista do consumidor?

DIFÍCIL – Exige o reconhecimento da argumentação, a apreensão de julgamento de valor além da recuperação da informação.

4.4.2 Aplicação do teste

Após a re-elaboração, o teste foi aplicado em 120 sujeitos, em diferentes instituições, com a seguinte distribuição de sujeitos por instituição:

UFPR – 47

UTFPR – 31

FACINTER – 15

CELIN – 12

UNIBRASIL – 11

TUIUTI – 04

Devido a questões de ordem prática (dificuldade de concentrar todos os sujeitos da pesquisa em um mesmo dia, hora e local para a aplicação do teste), os testes foram aplicados durante o período de aula dos sujeitos, em suas respectivas salas de aula, sob a supervisão dos seus professores.

O tempo de aplicação do teste (para este e para os testes-piloto) foi estabelecido, com base no tempo de teste estabelecido pela UFPR, instituição cujas diretrizes a autora está tomando por base para o desenvolvimento da pesquisa. O tempo dessas provas de suficiência é, em geral, duas horas e meia (2h e 30min). Entretanto, em função da dificuldade apresentada acima, o tempo de prova foi reduzido para uma hora e quarenta (1h e 40 min), tempo correspondente a duas aulas de 50 min. Verificou-se na prática, antes da

aplicação dos testes-piloto, que este tempo seria suficiente para a realização do teste, por meio de um pré-teste, em um público bem reduzido, embora com características semelhantes ao pretendido. Estes testes foram corrigidos apenas pela autora, com a finalidade de observar a real possibilidade de redução de tempo e não fizeram parte da pesquisa para a análise de dados.

As instruções de realização do teste, além das normalmente estabelecidas para a realização de testes em geral, foram:

a) duração máxima do teste = 1:40h;
b) utilizar caneta azul ou preta para responder as questões;
c) as questões são todas discursivas e escritas em português;
d) as respostas devem ser dadas em português;
e) é permitido o uso de dicionário durante a realização do teste;

Sendo assim, depois da escolha de um dos testes para re-elaboração e de sua posterior aplicação, conforme descrito acima, tendo em vista a metodologia adotada para esta pesquisa, para a análise dos resultados por meio de aplicativo de análise de dados, no caso, o Modelo Rasch, passou-se ao procedimento de correção, de acordo com a elaboração de descritores e treinamentos dos corretores e à análise dos resultados, conforme detalhado no capítulo a seguir.

5 CORREÇÃO DOS TESTES E ANÁLISE DOS RESULTADOS

Este capítulo visa apresentar os resultados da correção e da análise dos testes. Primeiro, será apresentado o processo de correção do teste final, relativamente aos procedimentos, tais como elaboração dos descritores e treinamento dos corretores, que foram adotados apenas para o teste final. A decisão de não adotar os mesmos procedimentos para os testes-piloto foi devida à intenção de que os resultados dos dois processos de correção pudessem ser comparados na análise dos gráficos. Segundo, será apresentada a observação e análise objetiva dos resultados obtidos no Modelo Rasch, que transforma a escala ordinal construída com os escores brutos atribuídos pelos corretores em uma escala intervalar, na qual se pode observar o desempenho dos indivíduos em termos de probabilidade de acerto das questões, o desempenho de cada grupo em relação aos outros dois em cada questão e a dificuldade das questões, em função do desempenho dos indivíduos em cada uma delas, por exemplo. A análise desses resultados obtidos será realizada, primeiramente apenas em relação aos resultados obtidos com a aplicação do piloto 1B, em cuja correção não foram estabelecidos critérios em comum entre os corretores e não foi realizado treinamento algum, conforme observado acima. Depois, a análise de resultados será realizada estabelecendo uma comparação entre esses resultados e os do teste final, que foi aplicado após alterações feitas em três questões do 1B (questões 3; 7 e 8), conforme explicado no item 4.4.1 do capítulo anterior e seguindo todos os procedimentos de correção apresentados no item 5.1, a seguir.

Como mencionado na revisão bibliográfica, alguns conceitos não foram apresentados naquele momento, para que fossem apresentados neste capítulo. O objetivo desse deslocamento de conceitos teóricos se deu em função da necessidade específica de se fazer uma relação direta entre eles e as análises dos dados. A apresentação desses conceitos imediatamente antes da análise evita que o leitor tenha que retornar ao terceiro capítulo em busca

dessas informações, antes de prosseguir a leitura, em determinados momentos.

5.1 PROCESSO DE CORREÇÃO DO TESTE FINAL

Como mencionado anteriormente, para a correção dos testes-piloto não foram estabelecidos descritores e não houve treinamento de corretores. A diferença de correção em função disso pode ser percebida, posteriormente conforme previsto, na análise da classificação do poder de adequação dos dois testes e na análise das questões. Sendo assim, o que será apresentado a seguir, diz respeito à correção do teste final.

Para a elaboração dos descritores utilizados na correção do teste final foi considerada a posterior utilização do modelo Rasch na análise dos dados. Neste modelo, a análise dos testes em questão é classificada como uma análise *politômica*[125] de dados, o que significa que para cada item analisado estarão envolvidas diferentes categorias ordenadas de resposta, com pontuação diferente para cada categoria.

Bachman e Palmer (1996, p. 212) afirmam que o número de níveis a ser estabelecido na elaboração dos descritores depende primeiramente da utilidade de cada nível. “Nós precisamos considerar o número de distinções que se pode sensatamente esperar que os corretores façam com confiabilidade e validade” (tradução nossa)[126]. Bond e Fox (2007, p. 221) afirmam ainda que, “o fato é, não há número ideal definitivo de categorias de resposta que se apliquem a todas as escalas de classificação. Enquanto cinco categorias de respostas podem funcionar para medir precisamente um construto, uma

[125] Na análise *dicotômica* de dados pressupõe o envolvimento de apenas duas categorias de resposta. Ou seja, é uma questão de *sim* ou *não*, de *certo* ou *errado*. Em geral estabelece-se que ou a resposta está errada e sua pontuação é 0 ou está certa e sua pontuação é 1.

[126] “We need to consider the number of distinctions raters can reasonably be expected do make reliably and validly.”

simples do tipo de resposta sim-ou-não pode ser melhor para outro." (tradução da autora)[127]

Considerando o exposto acima e o número de categorias do teste-piloto 1B, que apresentou uma dispersão grande dos respondentes, em função do número de sujeitos, foi construída uma escala de classificação para o teste final com 6 categorias, ou níveis de resposta, conforme apresentado a seguir.

5.1.1 Escala de Classificação e Descritores

A autora considerou o nível mais baixo de habilidade em termos de ausência de compreensão de leitura evidente e o mais alto em termos de proficiência na habilidade avaliada. Assim sendo a escala de classificação apresenta-se conforme abaixo.

Escala de classificação ordinal:

Categoria	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6
Pontuação	0	1	2	3	4	5

TABELA 6 – ESCALA DE CLASSIFICAÇÃO ORDINAL
FONTE: AUTOR (2008)

Na elaboração dos descritores dos níveis acima foram consideradas questões como: o construto de leitura a ser avaliado, já apresentado anteriormente; a necessidade de utilização de uma escala holística, em função da decisão de se avaliar a habilidade global do leitor e a dispensabilidade de elaboração de escalas com descritores para cada questão, em função ainda da abordagem holística a ser utilizada. Em lugar de se elaborar descritores específicos para cada questão, optou-se por elaborar apenas uma escala de

[127] "The fact is, there is no definitive optimal number of response categories that applies to all rating scales. Whereas five response categories might work for accurately measuring one construct, a simple yes-or-no type of response might be best for another."

classificação geral e também uma chave de respostas para as questões, para ser usada como modelo na correção.

Sendo assim, ficou estabelecido que:

a) escala ordinal para correção do teste:
   — escala com 6 níveis de diferenciação possíveis do leitor, onde o nível “0” (zero) indica que a tarefa não foi cumprida sob nenhum aspecto e o nível “5” (cinco) indica o cumprimento da tarefa nos padrões esperados de um leitor ideal proposto;
b) descritores dos seis níveis da escala ordinal de avaliação:
   — 5 – informações completas; resposta clara e com redação bem articulada, de forma a não deixar dúvidas quanto à compreensão do texto lido;
   — 4 – informações completas; resposta bem sucinta, apenas com palavras-chave ou com as informações imprescindíveis, sem apresentar o desenvolvimento das idéias e suas relações em forma de um texto coeso; inclusão de informações desnecessárias; pequenos problemas textuais;
   — 3 – informações incompletas, mas contendo no mínimo 70% da resposta esperada, sem problemas na sua redação; ou respostas completas, porém apresentando alguns problemas textuais, no desenvolvimento das idéias e/ou coesão entre elas; ou ainda apresentando a inclusão de alguma informação incorreta;
   — 2 – informações incompletas, contendo menos de 70% e mais de 40% da resposta esperada, com a inclusão de informações incorretas; com vários problemas textuais, no desenvolvimento das idéias e de coesão;
   — 1 – informações bastante incompletas (menos de 40% da resposta esperada); informações confusas; informações erradas; falta de textualidade na resposta, ou seja, resposta sem coesão e coerência, dificultando a compreensão da mesma; informações parcialmente ilegíveis;

— 0 – resposta em branco; nenhuma informação pertencente ou relacionada à questão; texto completamente ininteligível, em função de letra ilegível ou incoerência textual.

c) Linha de corte dos resultados do teste:

— em função da necessidade de se estabelecer um nível de conhecimento mínimo necessário aos estudos realizados nos cursos de mestrado e doutorado, foi estabelecido um nível, entre os seis apresentados acima, cuja habilidade descrita correspondesse ao critério estabelecido para o teste. Sendo assim, estabeleceu-se o nível 3 desta escala ordinal como sendo o nível mínimo exigido, ou nível de corte. Isso significa que, no caso de um teste real e não para fins de pesquisa como este, seria esperado que os candidatos atingissem no mínimo o nível 3, nessa escala ordinal, para que sua habilidade de leitura pudesse ser considerada suficiente de acordo com os propósitos do teste. No entanto, é importante salientar que o nível 3 desta escala ordinal não tem correspondência direta com o nível 3 de uma escala intervalar;[128]

d) Chave de respostas:

— a chave de respostas foi construída com as respostas esperadas atribuídas a um “leitor ideal”. O propósito desta chave é dispensar a necessidade de elaboração de descritores específicos para cada questão e fornecer aos corretores um ponto de partida para a correção de cada questão; uma possível resposta de valor 5, isto é, de nível 6 segundo a escala ordinal acima.

Em cada resposta esperada estão sublinhadas as informações consideradas indispensáveis à completude de cada uma delas.

---

[128] Essa correlação não será realizada neste trabalho, visto que estabelecer a linha de corte exata na escala intervalar não é relevante neste momento, por não contribuir nem interferir nas análises pretendidas e nas observações de dados propostas. No entanto, em um teste real, esta correlação precisaria ser feita.

QUESTÃO 1: Qual é o objetivo da pesquisa?

RESPOSTA ESPERADA: O objetivo da pesquisa é apresentar prós e contras da (produção da) proteína de insetos como nova fonte protéica, possível alternativa para a produção de carne.

QUESTÃO 2: O que motivou a pesquisa? Explique.

RESPOSTA ESPERADA: A motivação da pesquisa foi a necessidade de encontrar novas fontes de proteínas, como alternativa para a carne, tendo em vista o crescimento e o bem-estar da população mundial.

QUESTÃO 3: De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, escreva abaixo:

a) a denominação das 3(três) proteínas que foram apresentadas ao mercado consumidor;

RESPOSTA ESPERADA:
Tropina
Pruteen
Mycoprotein/Quorn

b) a reação do mercado consumidor (se for apresentada) e a razão do sucesso ou fracasso de cada uma delas.

RESPOSTA ESPERADA:

Tropina – o mercado consumidor questionou a condição de segurança para consumo do produto - foi um fracasso comercial, porque o preço do produto não conseguiu competir com o preço da soja, devido ao aumento do óleo. Pruteen – fracassou comercialmente porque o preço do produto não conseguiu competir com fontes alternativas de protéica como a soja e o peixe, quando houve aumento do preço do óleo.

Mycoprotein/Quorn – obteve mais sucesso que as proteínas anteriores - compõem produtos saudáveis, fáceis de preparar e que são similares às comidas conhecidas

QUESTÃO 4: Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento.

RESPOSTA ESPERADA: Embora o interesse pelos insetos como comida esteja começando a crescer no mundo ocidental, eles ainda não são considerados parte do grupo de alimentos. Eles são vistos com certo cepticismo e nojo.

Em vários países, no entanto, eles são aceitos e bastante consumidos em seus diferentes estágios de desenvolvimento.

QUESTÃO 5: Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína?

RESPOSTA ESPERADA: As espécies analisadas fornecem proteínas de alta qualidade e em quantidades equivalentes às encontradas na carne. Também suplementam a dieta significativamente com minerais e vitaminas. Além disso, a maioria dos insetos contém aminoácidos suficientes para suprir as necessidades nutricionais diárias.

QUESTÃO 6: Quais as vantagens e desvantagens da utilização da nova proteína como alimento?

Vantagens:

RESPOSTA ESPERADA: As vantagens mencionadas são:

- alta qualidade das proteínas e minerais e vitaminas que suplementam a dieta significativamente; (ou: além das vantagens apresentadas na questão anterior):

- a proteína do inseto é de fácil digestão;

- os insetos variam na quantidade de gordura e conseqüentemente na quantidade de energia fornecida;

- o valor calórico dos insetos varia dependendo da sua espécie e dieta;

- os ácidos graxos dos insetos são semelhantes aos encontrados no frango e no peixe;

- alguns insetos contêm ácidos graxos essenciais em maior quantidade que a carne;

- eles também apresentam alta concentração de minerais (zinco e ferro) e de cálcio.

Desvantagens:

RESPOSTA ESPERADA: As desvantagens mencionadas são:

- a composição e, portanto, o valor nutricional dos insetos varia;

- o aumento da criação de inseto pode ser difícil, porque nem todas as espécies são apropriadas para criação em larga escala;

- alguns problemas como vulnerabilidade às doenças e falta de cuidados recebidos podem surgir na criação;

- os insetos podem produzir secreções defensivas que podem causar alergias em quem os manuseia.

QUESTÃO 7: Diante da dificuldade que os consumidores demonstram em incluir os insetos em seus cardápios, os pesquisadores pensam em uma outra possibilidade. Diga qual a possibilidade e explique por que eles a consideram viável.

RESPOSTA ESPERADA: Em lugar dos insetos inteiros, pesquisadores pensam na possibilidade de utilizar células de insetos, cultivadas em suspensão em bioreatores.

A viabilidade dessa cultura deve-se ao fato de que o produto pode ser reproduzido sempre com a mesma qualidade; há um baixo risco de contaminação neste sistema, comparado com a criação de insetos; assim

como os insetos, as células não permitem o crescimento de vírus ou genes cancerosos que possam afetar seres humanos.

QUESTÃO 8: Que argumentos podem justificar mais estudos em relação ao ponto de vista do consumidor?

RESPOSTA ESPERADA: Do ponto de vista do mercado consumidor, para que uma nova proteína seja aceita, ela deve satisfazer os desejos desse consumidor. Os produtos devem ser, por exemplo, convenientes (de fácil consumo), saudáveis e similares a alguma outra comida que já seja familiar. No caso dos países ocidentais, deve-se ter em mente a percepção do consumidor em relação aos insetos e a possibilidade de essa percepção poder sofrer algum tipo de influência.

5.1.2 Treinamento de Corretores

Com base no exposto anteriormente sobre o treinamento de corretores o treinamento para a correção do teste final deu-se da seguinte forma:

a) a autora, elaboradora do teste foi a elaboradora da escala de classificação que foi utilizada pelos corretores;
b) foi elaborada uma escala com valores numéricos, para cada um dos quais foi elaborado um descritor. Além disso, foi elaborada uma chave de respostas para o teste, para estabelecer as informações desejadas pertinentes ao nível mais alto da escala, para minimizar a diferença de interpretação dos valores e do conteúdo das respostas entre os corretores;
c) foi utilizada uma escala de seis (06) níveis de pontuação, pois as distinções estabelecidas entre eles foram consideradas pela autora, e posteriormente pelos corretores, suficientes para os propósitos do teste;

d) embora o número de corretores tenha sido reduzido (04), estabeleceu-se a autora como 'o corretor chefe' (C);
e) após a aplicação do teste o corretor C leu alguns testes para levantar os tipos de respostas e problemas que elas apresentavam.
f) o corretor C extraiu algumas respostas que pudessem ser consideradas 'adequadas' e 'inadequadas' e problemas ainda não considerados;
g) o corretor C fez as adaptações necessárias nos descritores da escala de classificação;
h) em função do reduzido número de corretores, foi realizada apenas uma reunião de padronização da correção, da qual os quatro corretores participaram;
i) Todos os corretores receberam e usaram a escala de classificação para estabelecer os padrões de correção, compararam suas pontuações, discutiram as diferenças de opinião e estabeleceram, de forma consensual, uma pontuação para as amostras analisadas;
j) O treinamento demonstrou não haver necessidade de mudanças significativas na escala. Pequenas inclusões foram realizadas, resultando na escala apresentada acima.

5.1.3 Processo de Análise da Correção – O Modelo Rasch (TRI)

Depois de realizada a correção de todos os testes, foi elaborada uma tabela[129] contendo todos os dados a serem inseridos no aplicativo. Para que isso fosse possível, foram atribuídos números de 1 a 120 aos indivíduos que responderam os testes. A numeração foi atribuída, apenas por questões de organização, de acordo com a instituição onde o teste foi realizado, começando pela de maior número de testes realizados para a de menor número, e por ordem alfabética da primeira letra dos nomes dos sujeitos da pesquisa, sem levar em consideração nenhuma pontuação.

[129] Ver exemplo da tabela, referente a um grupo de dados de 20 testes no apêndice 4.

Os dados fornecidos pela correção necessários para rodar o aplicativo, neste caso, foram:

Identificação do Indivíduo	Identificação do Corretor	Pontuação das 8 questões
001	1	21342230
001	3	22310231

TABELA 6 – DADOS FORNECIDOS PELA CORREÇÃO
FONTE: AUTOR (2008)

A pontuação de cada questão foi dada por um dos níveis da escala de classificação, ou seja, por um único número, variando de 0 a 5 e foi colocada na tabela de forma seqüencial, começando pela 1ª questão, indo até a 8ª.

A análise dos resultados dos testes foi feita tomando por base os gráficos gerados a partir dos resultados submetidos à análise do modelo Rasch, conforme explicado brevemente acima. Não foi feita divisão entre a análise baseada na bibliografia e no modelo, entretanto ambos os recursos foram utilizados na medida em que foram necessários para tornar a análise o mais completa possível.

É importante salientar que não houve a intenção de trabalhar com todas as informações geradas pelo modelo, tanto pelo grande número de informações quanto por implicar na necessidade de apresentação e desenvolvimento de conhecimentos teóricos bastante específicos sobre estatística e análise, em vários desses dados, que não seriam relevantes para os objetivos pretendidos neste momento

Deve-se levar em consideração que a TRI, utilizada no Modelo Rasch, realiza a análise de dados em função da probabilidade de acerto de cada questão, com base nas informações fornecidas pelos resultados dos testes (como as apresentadas na tabela acima) e trabalha com dois axiomas fundamentais: 1) o desempenho do indivíduo em uma tarefa, onde o desempenho é o efeito da habilidade do indivíduo, que é explicada pelos parâmetros estabelecidos para o conjunto de traços latentes a serem observados em um item do teste; 2) a probabilidade de acerto que o indivíduo tem em dada tarefa (item do teste), que é dada pela relação entre o desempenho nesta tarefa e o conjunto de traços latentes. Essa relação é

descrita por uma equação monotônica crescente, denominada Função Característica do Item ou Curva Característica do Item (CCI).

A intenção da análise com base nos gráficos é apresentar o conjunto de dados que permita a verificação da validade do teste elaborado.

5.1.4 Curva Característica do Item (CCI)

Para que seja possível compreender a análise dos gráficos, é preciso antes entender o que é e como funciona a Curva Característica do Item (que de agora em diante será referida apenas como CCI). Para que isso seja possível, é preciso retomar, desenvolver, ou introduzir novos conceitos, que podem ter sido vistos (ou não) no Capítulo 2, relativos à Teoria de Resposta ao Item (TRI).

a) o desempenho de um candidato em um item do teste é explicado a partir de um conjunto de traços latentes. Neste caso, considera-se que o desempenho do candidato é o efeito e que os traços latentes são a sua causa (como explicado acima);
b) a relação que existe entre o desempenho do candidato no item e o conjunto de traços latentes é descrita na TRI e, portanto, no modelo Rasch, como uma equação monotônica (invariável) crescente, que é uma função matemática, chamada de Curva Característica do Item;
c) O que a CCI demonstra é que os candidatos com maior aptidão têm maior probabilidade de acertar o item e vice-versa. O gráfico da figura 5.1 abaixo pode ser interpretado da seguinte maneira:
   — É apresentado aos candidatos um estímulo (e.g. um item de um teste) ao qual eles respondem;
   — A resposta observada de cada candidato é relacionada com o nível de seu traço latente (habilidade de leitura), que é inferido a partir da análise da resposta dada;

— Esta relação pode ser expressa através de uma equação matemática, para a qual podem ser usados diversos modelos, que vai descrever a forma de função que a relação assume.[130];
— Essa função que a relação assume é descrita, em um gráfico, em forma de uma curva, a CCI ou Função Característica do Item;
— Os candidatos são divididos em grupos, de acordo com a capacidade demonstrada na realização das tarefas propostas;
— Caso o número de candidatos seja grande o suficiente para possibilitar a divisão em um maior número de grupos, com mais distinção de níveis de habilidade, os outros possíveis grupos de candidatos ficam distribuídos nos intervalos entre estes três níveis demonstrados no gráfico da figura 5.1;
— O parâmetro da dificuldade do item corresponde ao ponto na escala de habilidade onde a probabilidade de resposta é 0,50. Quanto maior for a dificuldade do item, maior deve ser o nível de habilidade do indivíduo a ser avaliado, para que ele tenha chance de 50% de acertar o item. Em outras palavras, cada ponto no gráfico representa o escore médio (onde: $0 \leq x \geq 1$) no eixo vertical para o grupo de indivíduos com habilidade estimada entre ± 0.2 *logits*[131] da localização no eixo horizontal;
— Tem-se então, de acordo com a figura, que os candidatos cuja habilidade é 4 têm 50% de probabilidade de acerto; os candidatos cuja habilidade é 1 têm bem menos de 10% de probabilidade; os de habilidade 8 têm probabilidade de acerto de 100%;
— A localização dos grupos (os pontos coloridos no gráfico abaixo) não é sempre coincidente com a CCI, como nessa figura. A localização dos grupos só coincide com a CCI se a média de acerto do grupo coincidir com a probabilidade de acerto prevista

[130]O objetivo deste trabalho é informar de onde surgem e o que são os dados que se encontram nos gráficos obtidos pelo modelo Rasch e que serão analisados a seguir e não entrar no mérito da questão matemática envolvida nesse processo de análise de dados. Para maiores detalhes sobre os cálculos envolvidos ver Pasquali (2004).

[131] Ver explicação desse termo no item 5.1.5, abaixo.

pela curva para aquele grupo. Se a média do grupo estiver abaixo do previsto, o grupo será localizado abaixo da curva, demonstrando que o desempenho demonstrado está aquém do previsto e se a média estiver acima, sua localização será acima da curva, demonstrando que o desempenho do grupo está além do previsto;

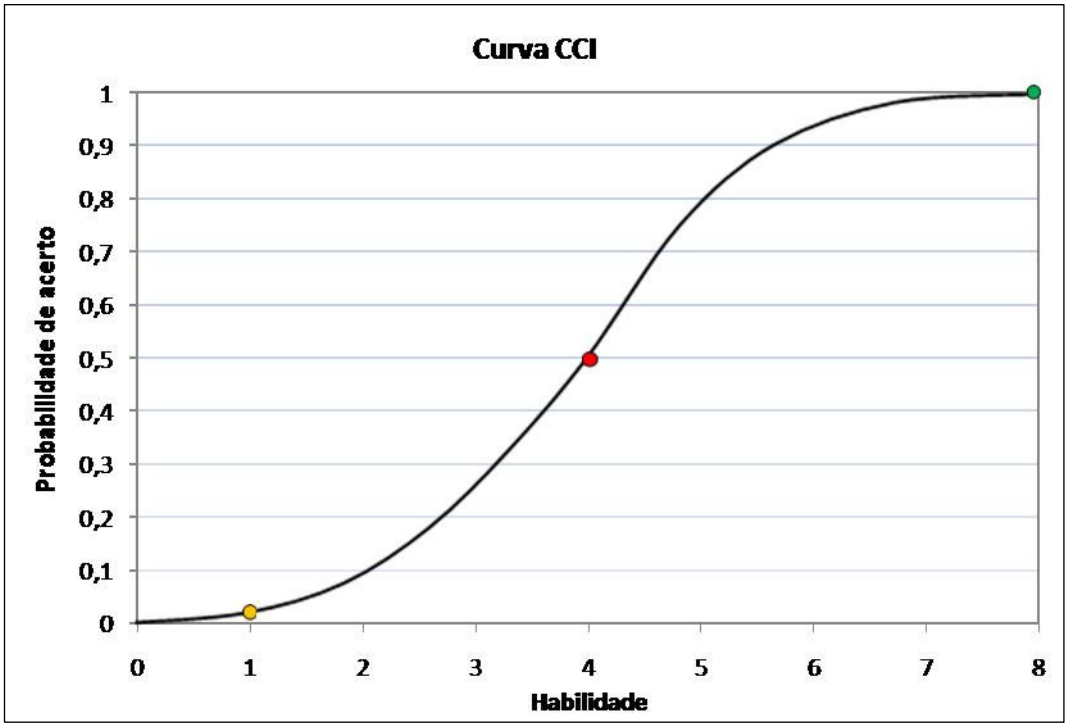


Figura 4 – CURVA CCI

FONTE: AUTOR – ADAPTADO DE PASQUALI (2008 – P.83)

Para explicar a matriz de dados que gera esse tipo de gráfico, Rasch (1960, p.117) apresenta um princípio que diz que

> uma pessoa que tem uma habilidade maior que outra pessoa deveria ter uma probabilidade maior de resolver qualquer item do tipo em questão, e da mesma forma, um item sendo mais difícil que o outro significa que para qualquer pessoa a probabilidade de resolver o segundo item é a maior."[132] (BOND AND FOX, 2007, p.10, tradução nossa).

[132]"a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one."

Essa matriz de dados é apresentada na forma de um quadro de probabilidades, como o da figura 5.2, e é construída para responder qual a probabilidade de uma pessoa com habilidade x (dado pelo número de itens respondidos corretamente) responder corretamente um item de dificuldade y (dado pelo número de pessoas que responderam este item corretamente). O que o modelo faz é usar um método de ordenação das pessoas de acordo com sua habilidade, e os itens de acordo com sua dificuldade. A partir disso, pode-se perceber que a probabilidade de alguém com uma habilidade acertar qualquer um dos itens, depende da diferença entre a habilidade da pessoa e a dificuldade do item. Essa matriz de dados é testada por axiomas matemáticos segundo os quais cada valor na tabela é sempre menor que: o valor a sua direita, o valor acima e o valor na diagonal à direita e acima.

TABLE 1.3
Table of Probabilities of Success When Ability Confronts Difficulty

Items / Persons	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*	Ability
P	.500	.866	.924	.949	.963	.973	.980	.986	.991	.995	.999	1/99
Q	.134	.500	.653	.741	.801	.847	.884	.915	.942	.968	.995	10/90
R	.076	.347	.500	.603	.682	.746	.801	.851	.896	.942	.991	20/80
S	.051	.259	.397	.500	.585	.659	.726	.789	.851	.915	.986	30/70
T	.037	.199	.318	.415	.500	.578	.653	.726	.801	.884	.980	40/60
U	.027	.153	.254	.341	.422	.500	.578	.659	.746	.847	.973	50/50
V	.020	.116	.199	.274	.347	.422	.500	.585	.682	.801	.963	60/40
W	.014	.085	.149	.211	.274	.341	.415	.500	.603	.741	.949	70/30
X	.009	.058	.104	.149	.199	.254	.318	.397	.500	.653	.924	80/20
Y	.005	.032	.058	.085	.116	.153	.199	.259	.347	.500	.866	90/10
Z	.001	.005	.009	.014	.020	.027	.037	.051	.076	.134	.500	99/1
Facility	1/99	10/90	20/80	30/70	40/60	50/50	60/40	70/30	80/20	90/10	99/1	

FIGURA 5 – TABLE OF PROBABILITIES OF SUCCESS WHEN ABILITY CONFRONTS DIFFICULTY

FONTE: BOND E FOX (2007 – P.11)

5.1.5 A escala intervalar

A dificuldade do item, por convenção, tem sua média ancorada em 0 (zero) e "se estende, na prática, de -3 a +3, passando por todas as decimais (de -3,00 até +3,00), obtendo +3,00 o item mais difícil e -3,00 o mais fácil." (PASQUALI, 2003, p. 130) As probabilidades de obtenção de uma resposta são expressas em termos de logaritmos (ou *log*) e as unidades de medida na escala construída pelo modelo são chamadas de *logits* (unidades logarítmicas de probabilidade – do inglês *log odds units*). Esta escala expressa em *logits* é uma escala intervalar e sua vantagem é fornecer a informação de **quanto um item é mais difícil que outro** e não apenas estabelecer a ordem de dificuldade entre eles. Convencionalmente então, a média de dificuldade dos itens de um teste é estabelecida em 0 (zero) *logits*. Isso significa que os itens que apresentam dificuldade acima da média são expressos por valores positivos e aqueles cuja dificuldade está abaixo da média estão expressos em valores negativos. As estimativas de habilidade dos indivíduos estão relacionadas à estimativa da dificuldade do item. Dessa forma, a habilidade média dos indivíduos também está ancorada em 0 (zero) *logits*, o que significa que um indivíduo cuja habilidade está localizada em 0 *logits* tem 50% de chance de acertar um item de dificuldade média. Assim como acontece com os itens, as habilidades acima da média são expressas por valores positivos e as abaixo da média por valores negativos.

"No Modelo Rasch, a mesma escala é usada para expressar a medida da habilidade da pessoa e a dificuldade dos itens, porque essas coisas são expressas em relação uma à outra; a escala logit é uma expressão da relação entre a dificuldade do item e a habilidade da pessoa." (MCNAMARA, 1996, p. 166-7, tradução nossa). Essa é a propriedade mais útil da análise Rasch, de acordo com este autor. Uma das razões que ele apresenta é a possibilidade de verificação se um grupo de itens é muito fácil ou muito difícil para um determinado grupo de indivíduos. Este mapeamento da dificuldade dos itens e da habilidade do grupo testado permite que se perceba quão apropriado é o teste para aquele grupo e até que ponto determinado grupo provavelmente se

sairia bem em determinado teste. Esse foi o tipo de verificação feita nos dados dos testes Piloto 1B e Final, apresentada a seguir.

5.2 ANÁLISE DOS RESULTADOS - PILOTO 1B x TESTE FINAL

A análise dos dados gerados pelo aplicativo foi realizada com base nos gráficos das curvas características dos itens e em algumas outras informações, referentes aos dois testes: piloto 1B e teste final. Foram analisadas, dentre todas as informações geradas, apenas aquelas consideradas relevantes aos objetivos propostos.

Na tabela que resume as estatísticas dos dados do teste, algumas informações, em especial, foram relevantes no momento de análise. Para melhor entendimento da análise dos gráficos, convém elucidar quais são as informações e o que elas representam. Para isso, reproduziu-se logo abaixo a tabela da análise do piloto 1B[133], com as informações destacadas, seguida das explicações relativas a cada uma delas.

a) *Item Location* (Localização do Item) – A média de localização dos itens é estabelecida em **0,0.** Esta localização é uma restrição arbitraria imposta por programas como o utilizado nesta análise, porque apenas a localização relativa do item pode ser estimada, e não a sua localização absoluta. Por definição, então, a escala é construída a partir do **0** (zero).

[133] Essa tabela também se encontra no apêndice 3, juntamente com a tabela relativa ao Teste Final, para comparação de dados.

```
RUMM2020        Project: PILOTO1BB        Analysis: PILOTO1B
Title:  PILOTO1B                        Date: 7 out 2008 01:33:41
Display: SUMMARY TEST-OF-FIT STATISTICS

                        ITEM-PERSON INTERACTION
=========================================================================
                          ITEMS                          PERSONS
                  Location  Fit Residual      Location  Fit Residual
-------------------------------------------------------------------------
Mean                 0,000       0,578         0,254       0,274
SD                   0,789       0,722         0,320       0,799
Skewness                         0,024                     0,425
Kurtosis                        -1,183                    -0,938
Correlation                     -0,015                     0,144

Complete data DF =               0,782
-------------------------------------------------------------------------

=========================================================================
        ITEM-TRAIT INTERACTION            RELIABILITY INDICES
-------------------------------------------------------------------------
Total Item Chi Squ           49,555     Separation Index 0,79341
Total Deg of Freedom         32,000     Cronbach Alpha      N/A
Total Chi Squ Prob          0,024592
-------------------------------------------------------------------------

=========================================================================
        LIKELIHOOD-RATIO TEST             POWER OF TEST-OF-FIT
-------------------------------------------------------------------------
Chi Squ                                  Power is GOOD
Degrees of Freedom              [Based on SepIndex of 0,79341]
Probability
-------------------------------------------------------------------------
```

FIGURA 6 – THE SUMMARY STATISTICS
FONTE: AUTOR (2008)

b) *Item - Fit Residual* (Resíduo) – Esta é uma estatística que fornece a informação sobre a adequação dos dados ao modelo sob a perspectiva dos itens. Essa adequação é interpretada da seguinte forma: se os dados estão adequados ao modelo, a sua média relativa a todos os itens deve estar próxima de **0** e o Desvio Padrão (*Standar Deviation)* dessa adequação deve estar próxima de **1.** Essa informação pode ser apreendida através do gráfico também. Ela é dada pela distribuição dos pontos em relação à CCI. Quanto menor a distância entre o ponto (escore médio obtido por um grupo) e a curva, mais adequados os dados observados estão ao modelo; em outras palavras, o escore observado estará mais próximo do esperado. A adequação seria total, caso os escores fossem iguais. Neste caso, o resíduo seria

0 (zero) e o ponto estaria localizado no gráfico exata e precisamente em cima da curva.

Além dessas informações, é importante esclarecer que as informações levantadas pelo aplicativo são apresentadas de forma padronizada nos gráficos, como este abaixo. Elas podem ser lidas conforme explicação que se segue:

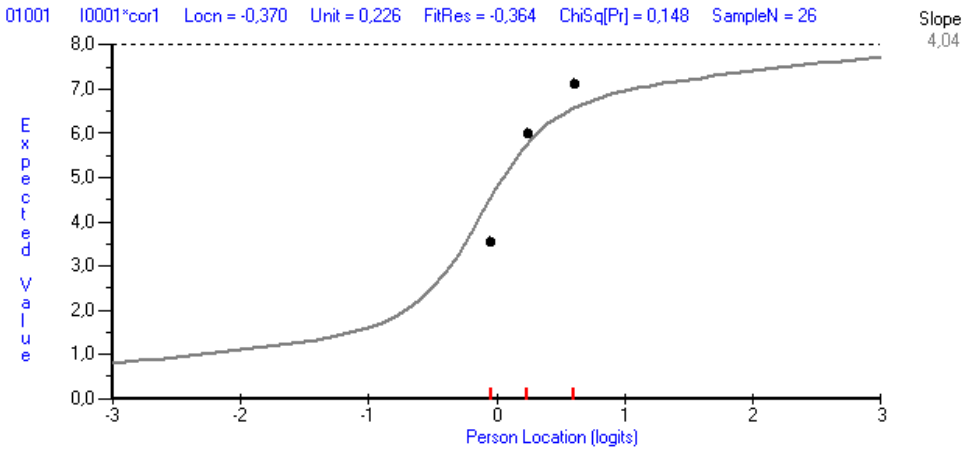


GRÁFICO 1 – QUESTÃO 1 DO PILOTO 1B – CORRETOR 1
FONTE: AUTOR (2008)

a) para a identificação dos gráficos é utilizada a segunda numeração acima à esquerda. No gráfico acima: **I0001*cor1**, que significa que este é o Item 1 (ou questão 1), corrigido pelo corretor 1;
b) na última informação à direita, acima do gráfico, onde se lê *SampleN*, o número 26 corresponde ao número de indivíduos que realizou esse teste;
c) considerando o número de dados fornecido, o programa fez a divisão dos indivíduos que realizaram os testes em três grupos, que estão classificados por ordem de desempenho. Dessa forma, existe um grupo mais fraco, com menores probabilidades de acerto; um grupo intermediário e um grupo mais forte, com maiores probabilidades de acerto das questões;
d) esses três grupos são representados por pontos, conforme explicado anteriormente, na explicação sobre a CCI;
e) o eixo vertical do gráfico refere-se à probabilidade de acerto que o indivíduo tem em determinada questão. E o horizontal à localização

do indivíduo, de acordo com a habilidade que ele demonstra ao responder aquele item (questão).

A partir dessas explicações se procederá a análise, com a inclusão de outras informações, que se façam necessárias conforme o caso.

Apenas para efeito das análises que se seguem, ficará estabelecido que:

a) para facilitar a localização dos grupos nos gráficos e a forma de se referir a eles será designada uma denominação para cada grupo. Sendo assim, o grupo mais hábil será denominado grupo A, o intermediário, grupo B e o menos hábil, grupo C;
b) os corretores serão designados por números, de acordo com a numeração dos gráficos (por exemplo: cor1, cor2,...).
c) a localização dos grupos em relação ao eixo horizontal é marcada por pequenos traços verticais;
d) para melhor visualização da análise, será apresentada a questão, ou questões, os gráficos a ela relacionados e os comentários pertinentes a cada caso, sempre nesta ordem.

5.2.1 Piloto 1B

Na análise do piloto 1B (doravante denominado simplesmente 1B), percebeu-se, como esperado, que em função do não estabelecimento de descritores específicos e únicos para serem utilizados pelos dois corretores e da falta de treinamento, houve diferenças de correção o suficiente para comprometer a confiabilidade de teste. Além disso, pode-se confirmar a dificuldade/facilidade de determinadas questões e observar quais delas se constituíram em problemas, comprometendo a validade do teste, em função de sua (má)formulação, justificando uma alteração das questões ou sua substituição para o teste final (doravante denominado simplesmente TF).

Questão 1 (1B): Qual é o objetivo da pesquisa?

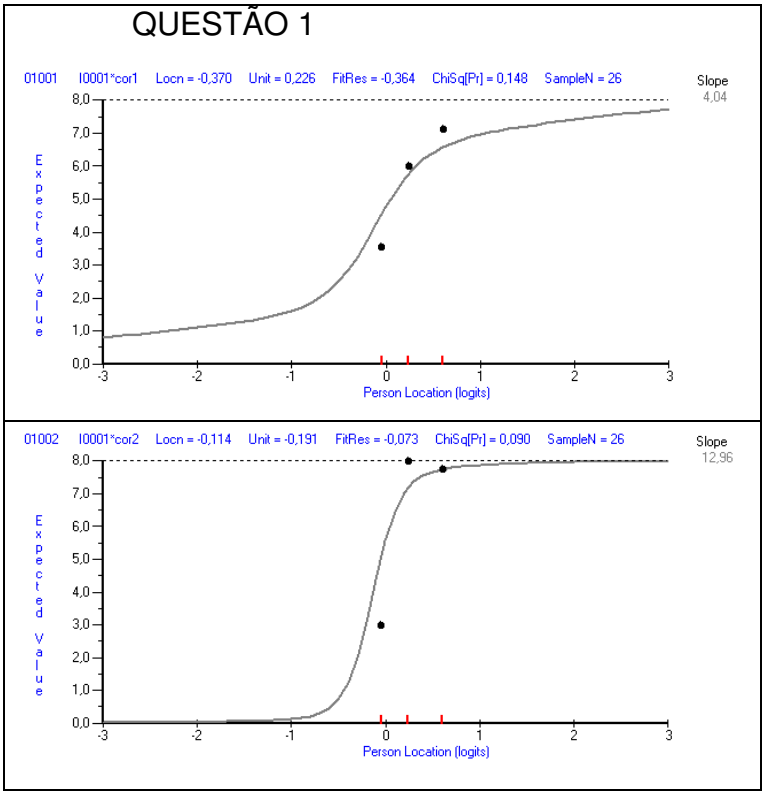


GRÁFICO 2 – QUESTÃO 1
FONTE: AUTOR (2008)

A análise desses dois gráficos demonstra que o problema com esta questão pode ter como fonte uma divergência de critérios de correção usados pelos dois corretores.

O gráfico do cor1 apresenta uma distribuição de resultados mais equilibrada de acordo com a Curva de Característica do Item. A localização dos grupos está distribuída ao longo da curva com cada grupo obtendo o desempenho bastante próximo do previsto, ou seja, o grupo C obteve os menores escores, o B os escores intermediários e o grupo A os escores mais altos, como previsto pela curva. Além disso, a localização dos grupos está razoavelmente próxima à CCI, mostrando não ter havido desvios significativos do padrão estabelecido.

Isso não acontece com o gráfico do cor2, que apresenta uma distribuição de dados na CCI com um desvio bastante acentuado. O que se observa é que o grupo B obteve um desempenho superior ao do grupo A, que em princípio deveria ter atingido a melhor média de escores; fato indicativo de alguma anormalidade por não estar de acordo com a curva prevista.

Questão 2 (1B): O que motivou a pesquisa? Explique.

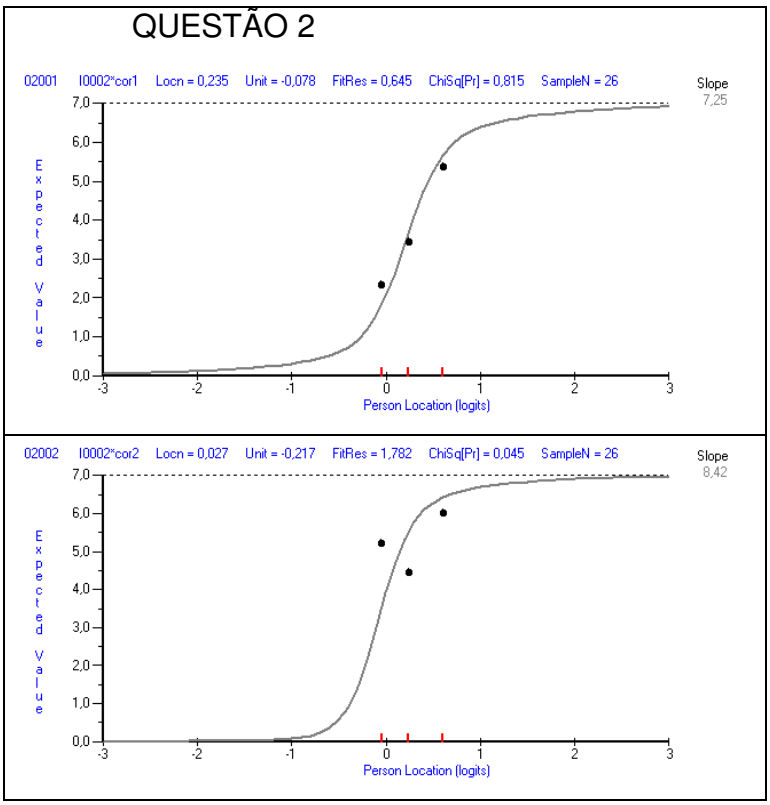


GRÁFICO 3 – QUESTÃO 2
FONTE: AUTOR (2008)

Nesta questão o cor2 foi mais leniente em sua correção que o cor1. Pode-se observar que todos os grupos estão com média de escores elevada. Além disso, o grupo C, para quem a questão deveria ser mais difícil, recebeu escores mais altos que o B, localizado um pouco acima de zero no eixo horizontal. Todos os grupos, diferentemente do observado para o cor1, estão afastados da curva, mostrando, ainda, que os escores estão bastante diferentes do previsto.

Questão 3 (1B): De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, preencha a tabela abaixo.

Proteína	Sucesso ou Fracasso	Razão

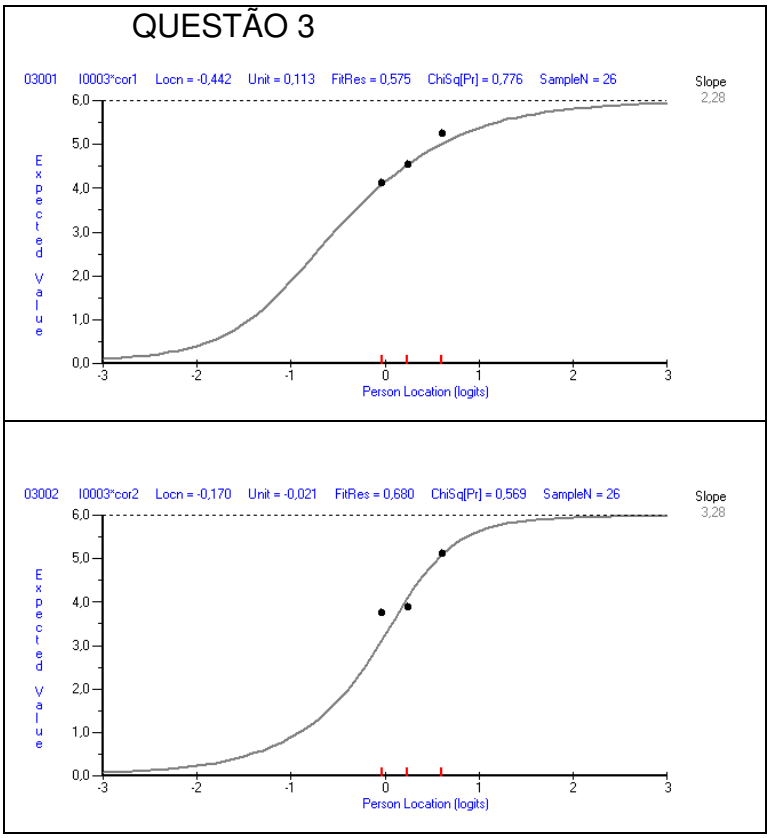


GRÁFICO 4 – QUESTÃO 3
FONTE: AUTOR (2008)

Esta foi uma questão muito fácil. Pode-se observar que os três grupos obtiveram escores acima da média para os dois corretores. Percebe-se, também, que não houve maiores divergência entre os corretores, ou problemas muito grandes de correção para qualquer um deles. Entretanto, o escore dos grupos está bastante próximo, mostrando que não houve grande diferença de desempenho entre um grupo e outro, fato que pode ser reputado à facilidade da questão. No caso do cor2, os grupos B e C estão praticamente com o mesmo escore.

Conforme levantado na correção e re-elaboração do 1B anteriormente, esta questão foi modificada por ser fácil em demasia, não contribuindo, dessa forma, para uma discriminação mais efetiva entre os leitores.

Questão 4 (1B): Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento?

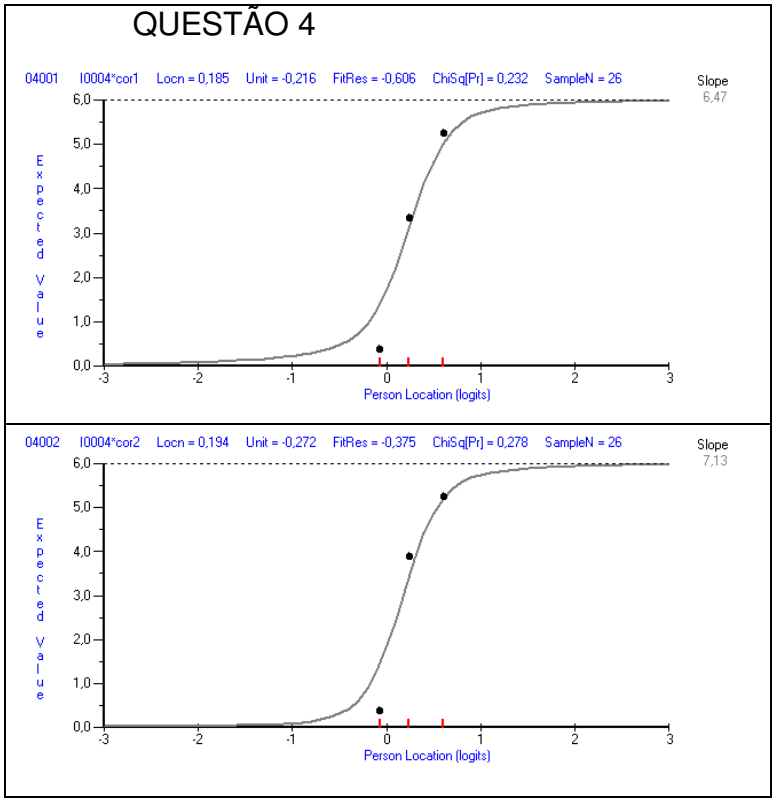


GRÁFICO 5 – QUESTÃO 4
FONTE: AUTOR (2008)

Esta foi uma questão considerada muito boa, pois proporcionou a discriminação entre os leitores e não apresentou grandes diferenças entre o desempenho esperado e o obtido, a não ser no grupo C. Teoricamente, esta é uma questão de dificuldade média. Pelos gráficos pode-se perceber que o grupo B demonstra ter probabilidade de acerto acima de 50%, mas ainda distante da probabilidade do grupo A. Isso confirma a análise da dificuldade da questão apresentada anteriormente e justifica a sua manutenção no TF.

Questão 5 (1B): Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína?

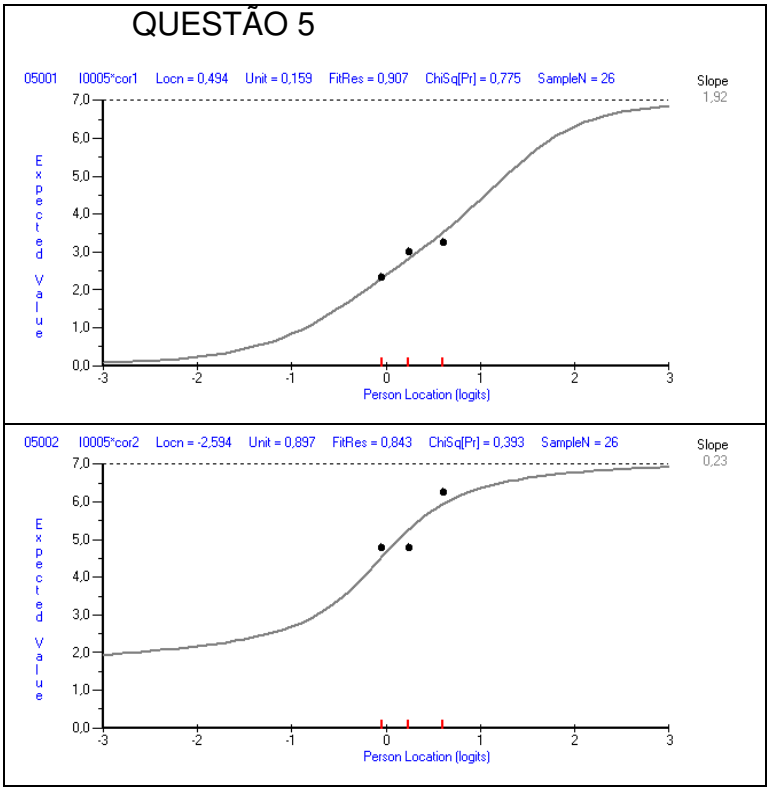


GRÁFICO 6 – QUESTÃO 5
FONTE: AUTOR (2008)

Esta questão foi classificada como fácil, no item 4.4.1 do capítulo anterior. Isso se mostrou verdadeiro para o cor2, cujo gráfico apresenta a localização dos grupos acima da média esperada. Observa-se que para este corretor o valor mínimo previsto pela curva está bem próximo de 2,0. No entanto, isso acontece diferentemente para o cor1, em cujo gráfico nota-se a alocação dos grupos um pouco abaixo da média, indicando que para este corretor esta questão é de dificuldade entre mediana e difícil, denotando uma diferença de rigor entre eles.

Outro problema que se verifica é que a diferença entre os grupos é muito pequena, mostrando que a questão não discriminou muito bem. Além disso, há também uma pequena anomalia no gráfico do cor2, em cujos grupos B e C receberam escores iguais.

Questão 6 (1B): Quais as vantagens e desvantagens da utilização da nova proteína como alimento?

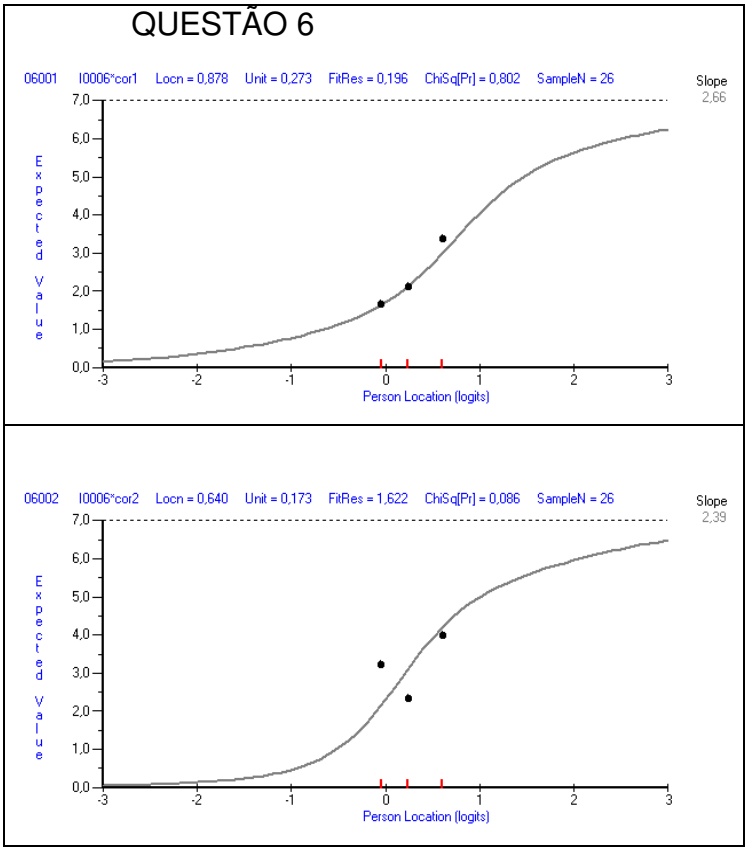


GRÁFICO 7 – QUESTÃO 6
FONTE: AUTOR (2008)

Observa-se nos gráficos que não houve diferença de rigor entre os corretores. No entanto o gráfico do cor2 apresenta uma anomalia, pelo fato de que o grupo C obteve desempenho melhor que o B.

Questão 7 (1B): Os autores do artigo chegaram a que conclusão?

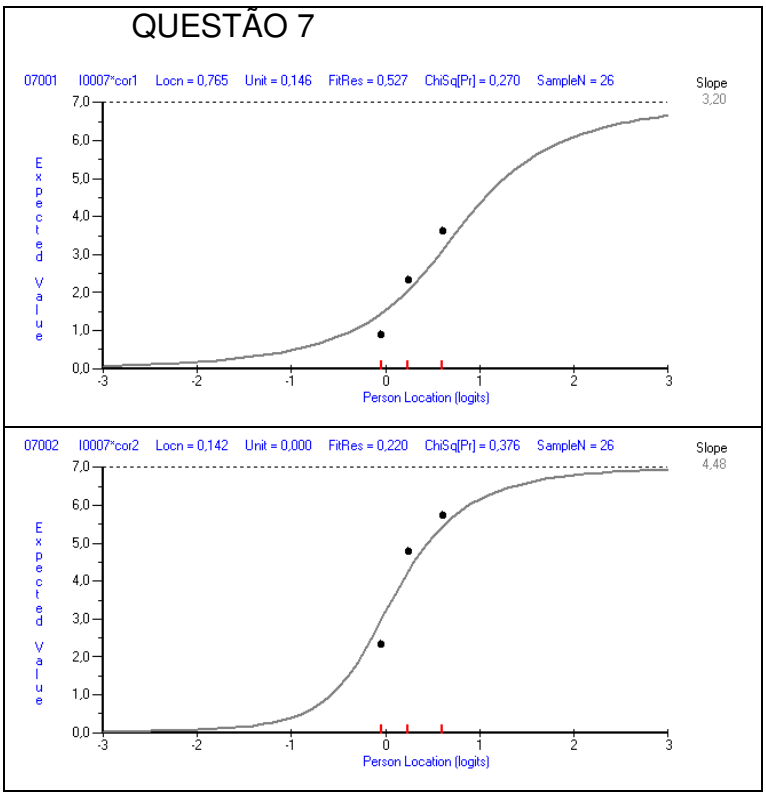


GRÁFICO 8 – QUESTÃO 7
FONTE:AUTOR (2008)

Nesta questão percebe-se claramente uma diferença de rigor entre os corretores, sendo que o cor1 foi o mais rigoroso. No entanto, o posicionamento dos grupos em relação uns aos outros e em relação à curva se mostra de forma semelhante nos dois corretores, denotando uma provável adoção de critérios de correção semelhantes nesta questão.

Questão 8 (1B): Qual é o papel do consumidor nessa história?

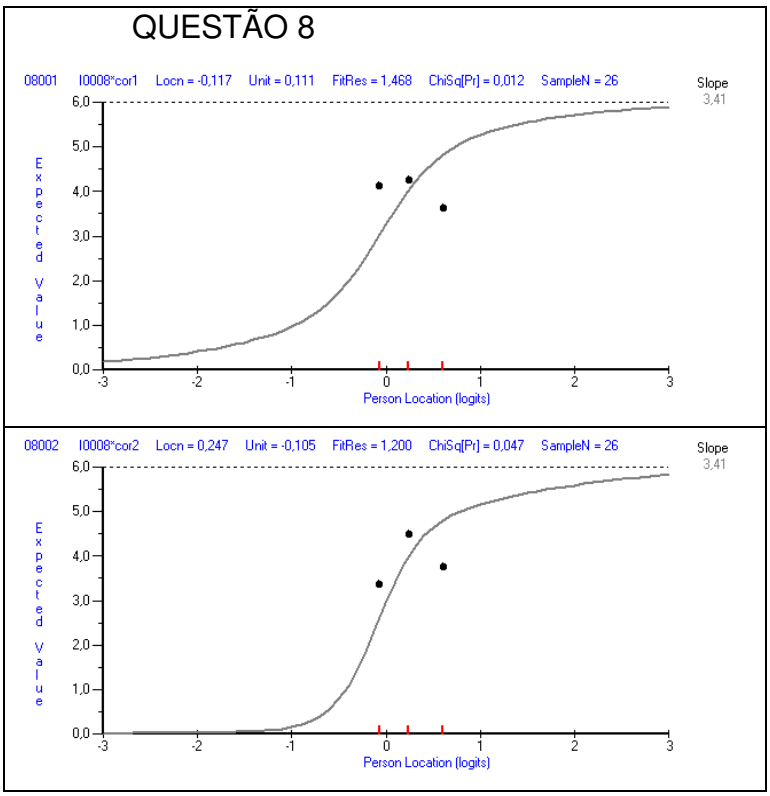


GRÁFICO 9 – QUESTÃO 8
FONTE:AUTOR (2008)

Esta questão não foi considerada uma boa questão. Os gráficos dos dois corretores mostram inconsistências em relação a todos os grupos, tanto na sua localização quanto na sua adequação à curva. Isso demonstra na prática a necessidade de substituir a questão, conforme observado durante a correção e relatado na re-elaboração do teste, no capítulo 4.

Problemas relacionados à elaboração da questão, foram identificados nos gráficos das questões de número 3 e 8, como previsto pela correção. Além dessas duas, a questão 7 havia sido considerada um tanto ambígua e, embora não tenha apresentado problemas muito significativos na análise dos dados, optou-se por alterá-la mesmo assim, na tentativa de obter melhores resultados.

Os problemas relatados acima, em relação às diferenças entre os corretores e as inconsistências de um ou de outro corretor, nas demais questões, foram atribuídas, em um primeiro momento, à falta de estabelecimento de critérios para uso comum e de treinamento entre os

corretores. Essa atribuição ficou comprovada, com a observação dos dados gerados pela correção do teste re-elaborado, o Teste Final, para cuja correção foram elaborados os descritores e realizado o treinamento, relatados anteriormente, no item 5.1. A análise desses dados será apresentada a seguir, em comparativamente aos dados do piloto 1B, que acabaram de ser vistos.

5.2.2 Piloto 1B X Teste Final

Tendo em mente a análise de cada questão realizada no 1B, será feita, a partir de agora, uma analise comparativa entre os resultados do TF (Teste Final) e os do 1B (piloto 1B), no que diz respeito àqueles problemas considerados mais significativos,[134] entre eles a falta de critérios em comum e o treinamento de corretores, ausentes no piloto 1B, mas presentes no TF.

É possível observar pela comparação dos gráficos de todas as questões, gerados para cada um dos quatro corretores, que o treinamento dos corretores e o uso dos descritores possibilitaram a realização uma correção bastante consistente entre os corretores. O cor4 é quem apresenta maiores diferenças em relação aos outros três, porém, não são diferenças significativas o bastante para colocar em risco a validade do treinamento e do uso dos descritores, como se pode observar nos gráficos abaixo.

[134] Para possibilitar uma comparação visual de todos os gráficos, dos dois testes, foi disponibilizada a reprodução de todos eles no apêndice 5.

Questão 1: Qual é o objetivo da pesquisa?

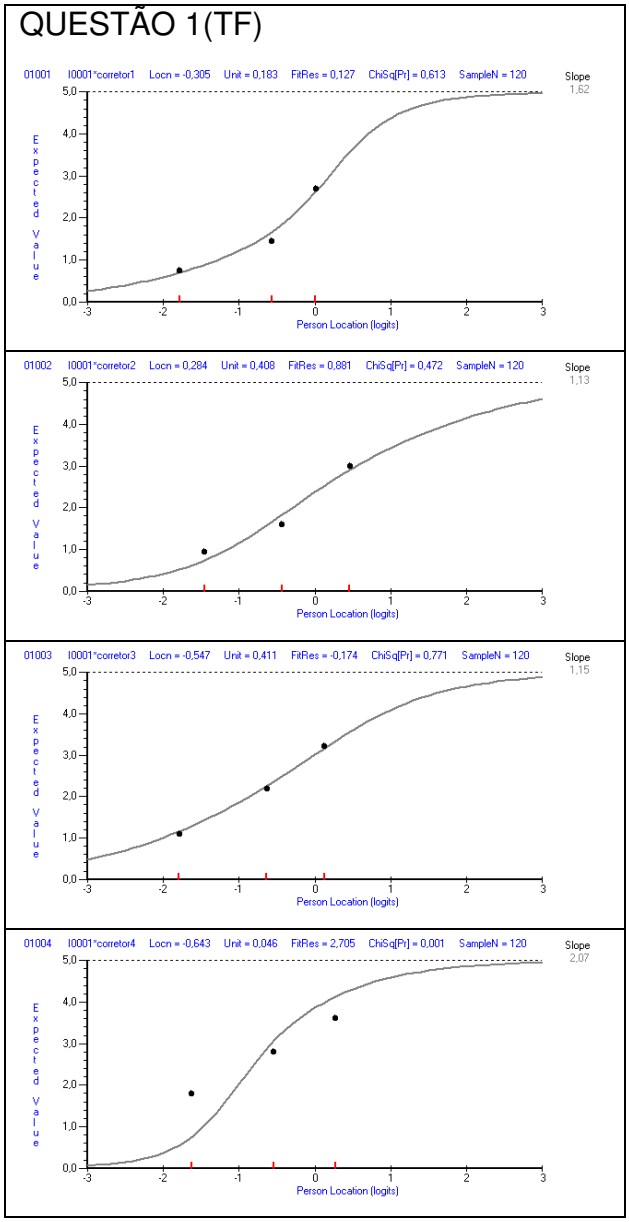


GRÁFICO 10 – QUESTÃO 1 (TF)
FONTE: AUTOR (2008)

Lembrando que esta questão não sofreu modificação alguma, pode-se observar que o gráfico dessa questão, para os quatro corretores, em relação ao teste 1B encontra-se significativamente modificado, com uma distribuição dos grupos bem mais adequada à CCI e também em relação à sua localização, no que concerne à habilidade de cada um. Se esses quatro gráficos forem comparados ao gráfico do cor1 do piloto 1B, será possível perceber que a forma como localização dos grupos no eixo da habilidade está distribuída mostra uma discriminação bem maior entre os grupos.

Questão 2: O que motivou a pesquisa? Explique.

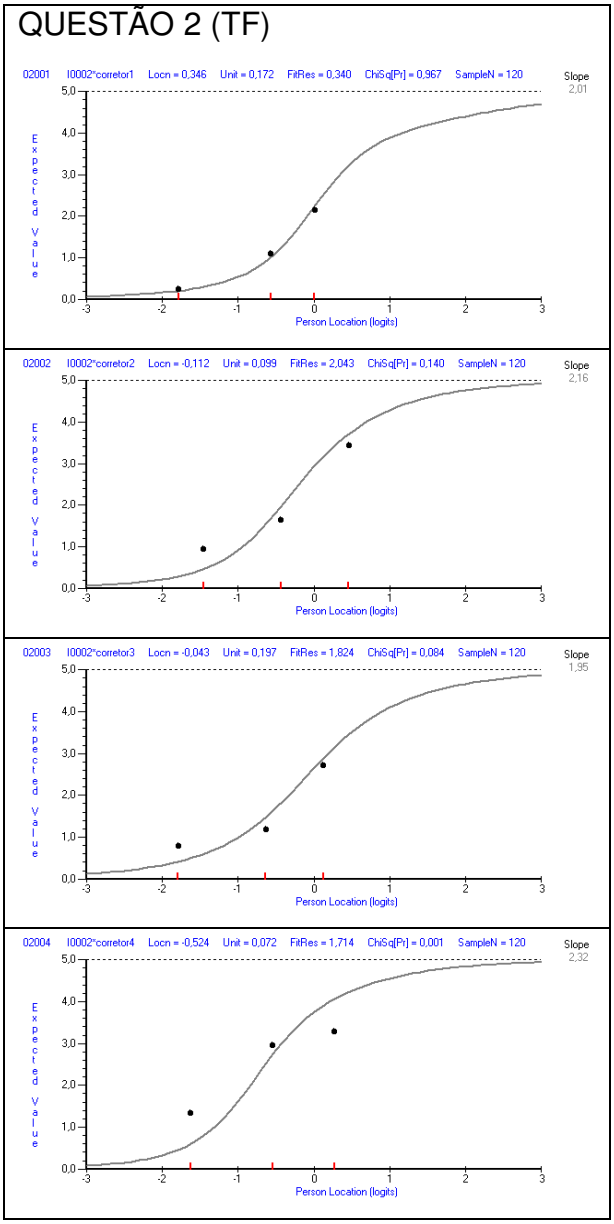


GRÁFICO 11 – QUESTÃO 2 (TF)
FONTE:AUTOR (2008)

A localização dos grupos nos gráficos do TF mostra que os critérios de correção e o treinamento levaram os corretores a adotar um grau de exigência maior em relação às respostas desta questão. No teste 1B, de maneira geral, os corretores foram mais lenientes que no TF. Nesta questão, o padrão dos grupos em relação à curva para os três primeiros corretores mostra uma correspondência bastante grande, embora o cor1 tenha sido o mais rigoroso e o cor2 o menos rigoroso dos três. Assim como acontece na questão 1, a discriminação entre os grupos apresenta-se mais acentuada.

Questão 3: De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, escreva abaixo:

a) a denominação das 3(três) proteínas que foram apresentadas ao mercado consumidor;
b) a reação do mercado consumidor (se for apresentada) e a razão do sucesso ou fracasso de cada uma delas.

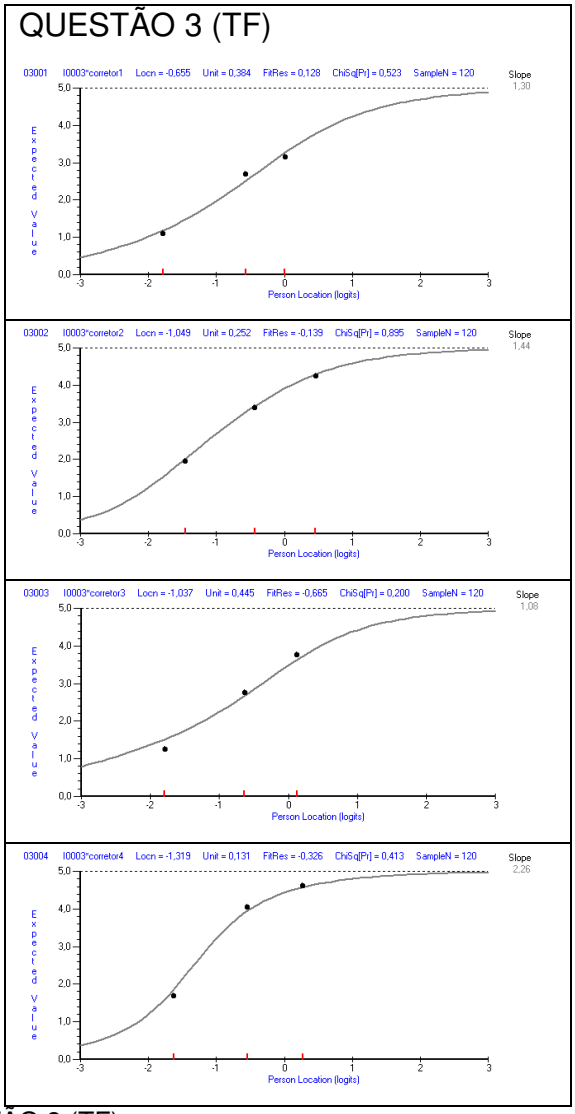


GRÁFICO 12 – QUESTÃO 3 (TF)
FONTE: AUTOR (2008)

A questão 3 foi re-elaborada por ter sido considerada demasiadamente fácil. A reformulação foi positiva, como se pode observar pelos gráficos acima e fez com que a questão ficasse um pouco mais difícil, conforme esperado. A discriminação entre os grupos também melhorou consideravelmente, em relação ao resultado anterior. Além disso, pode-se observar que o desempenho dos grupos, nesta questão, se ajusta quase que perfeitamente ao previsto pela CCI, independentemente do corretor observado (cor1, 2, 3 ou 4).

Questão 4: Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento.

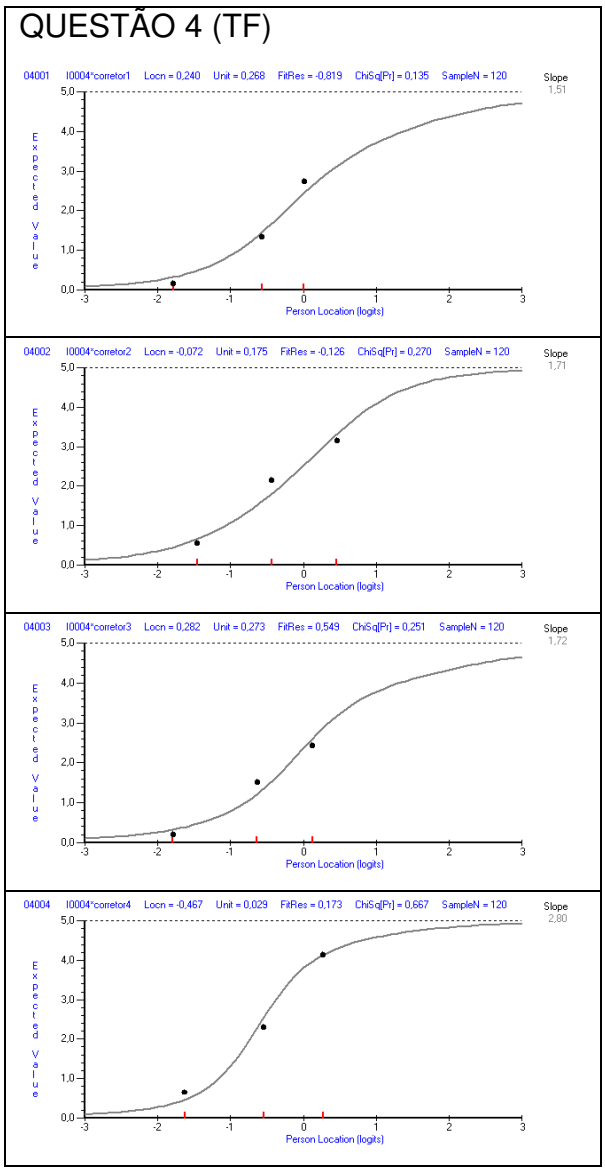


GRÁFICO 13 – QUESTÃO 4 (TF)

FONTE:AUTOR (2008)

Para os quatro corretores, esta questão mostrou-se um pouco mais difícil neste teste do que no piloto 1B. Essa diferença, provavelmente deve-se à adoção de critérios em comum e ao treinamento dos corretores. Isso fez com que o nível de rigor de todos eles ficasse um pouco mais elevado, diminuindo a diferença de valores dos escores dos grupos A e B, comparados aos escores desses grupos no piloto 1B.

A localização do nível de habilidade dos grupos, que no teste 1B estava bem concentrada próximo ao 0 no eixo horizontal, torna mais evidente a discriminação entre eles, no TF.

Questão 5: Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína?

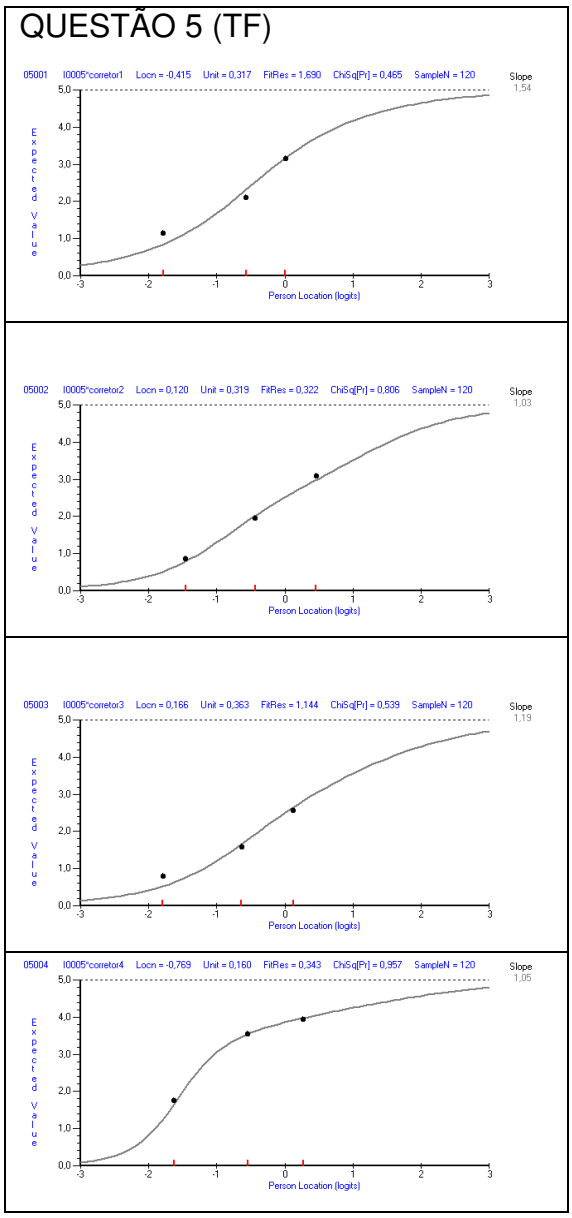


GRÁFICO 14 – QUESTÃO 5 (TF)

FONTE:AUTOR (2008)

Embora o gráfico do cor4 apresente-se um tanto diferente dos outros três, por ele ter sido um pouco mais leniente, existe uma uniformidade na distribuição dos grupos em relação ao eixo horizontal do gráfico, para todos os corretores. Comparando os gráficos dos dois testes, pode-se perceber que no TF esta distribuição está consideravelmente melhor que no teste 1B, cujos gráficos mostram uma concentração dos grupos muito próxima do 0 no eixo horizontal. Além disso, com o uso dos descritores, os escores no TF ficaram com valores bem mais discriminativos.

Questão 6: Quais as vantagens e desvantagens da utilização da nova proteína como alimento?

Vantagens:

Desvantagens:

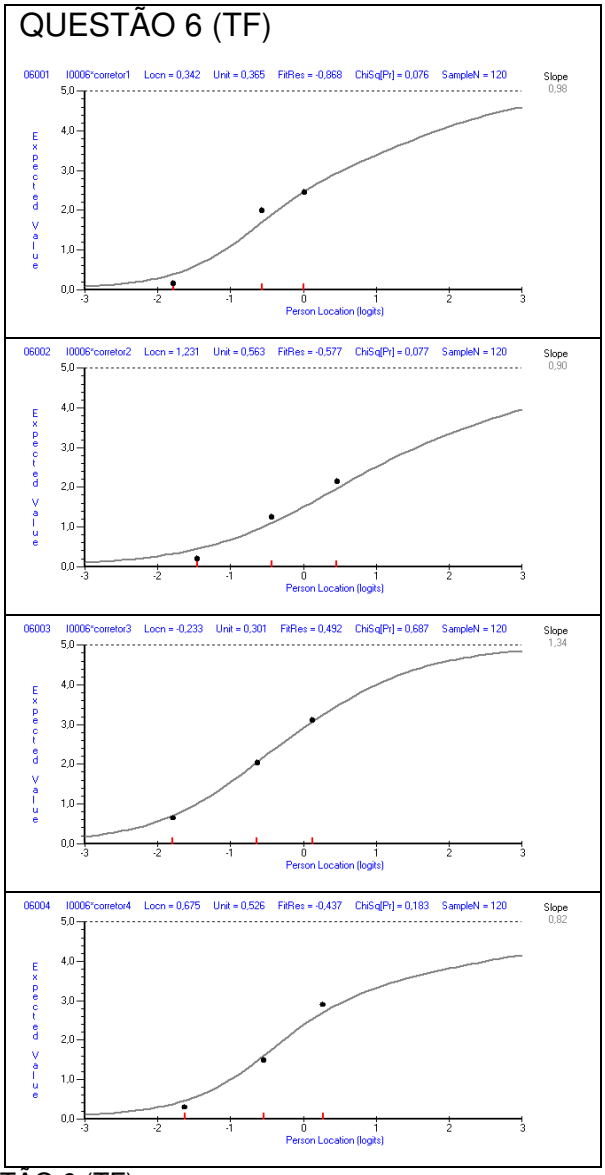


GRÁFICO 15 – QUESTÃO 6 (TF)
FONTE: AUTOR (2008)

Um dos gráficos relativos a esta questão no teste piloto apresentou uma anomalia que, pode-se perceber, não está presente em nenhum dos gráficos acima. Como os gráficos das outras questões, este também demonstra a consistência que o treinamento dos corretores proporcionou ao processo. Percebe-se que há pequenas variações entre os corretores em termos de rigor, porém elas são perfeitamente aceitáveis, visto que a subjetividade do processo não pode ser totalmente eliminada, apenas minimizada, como é o caso.

Questão 7: Diante da dificuldade que os consumidores demonstram em incluir os insetos em seus cardápios, os pesquisadores pensam em uma outra possibilidade. Diga qual a possibilidade e explique por que eles a consideram viável.

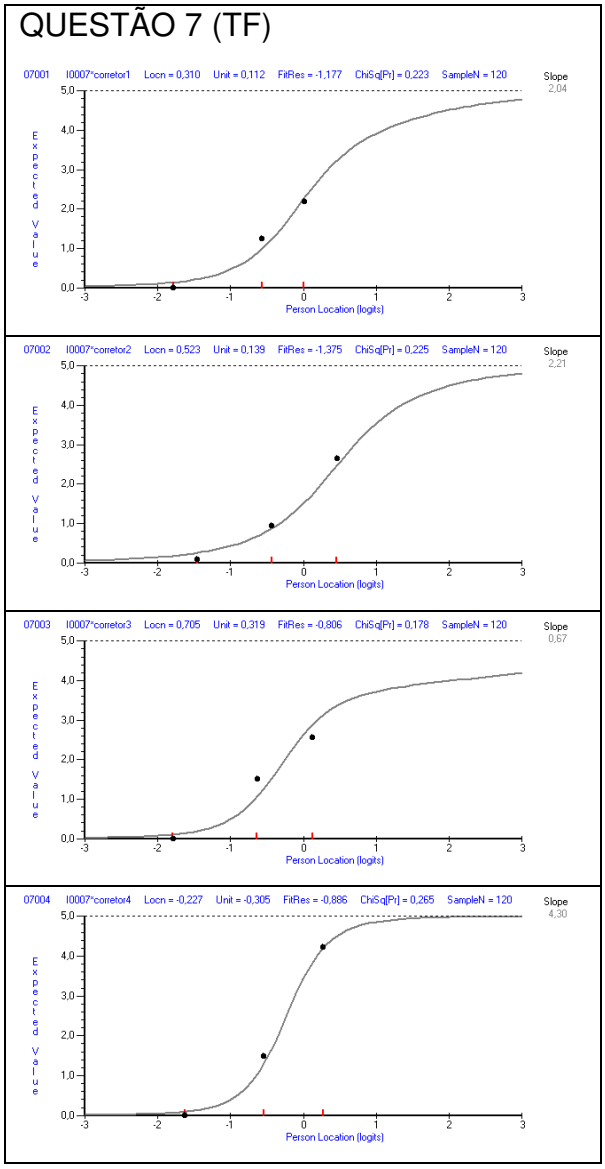


GRÁFICO 16 – QUESTÃO 7
FONTE: AUTOR (2008)

Essa foi a segunda questão alterada; na verdade ela foi substituída. A questão retirada não apresentava grandes problemas, conforme a análise dos gráficos anteriores. No entanto, pela análise dos gráficos acima, percebe-se que, apesar de algumas diferenças em termos de rigor entre os corretores, o desempenho dos candidatos pode ser avaliado com muito mais consistência, demonstrando que a substituição foi apropriada.

Questão 8: Que argumentos podem justificar mais estudos em relação ao ponto de vista do consumidor?

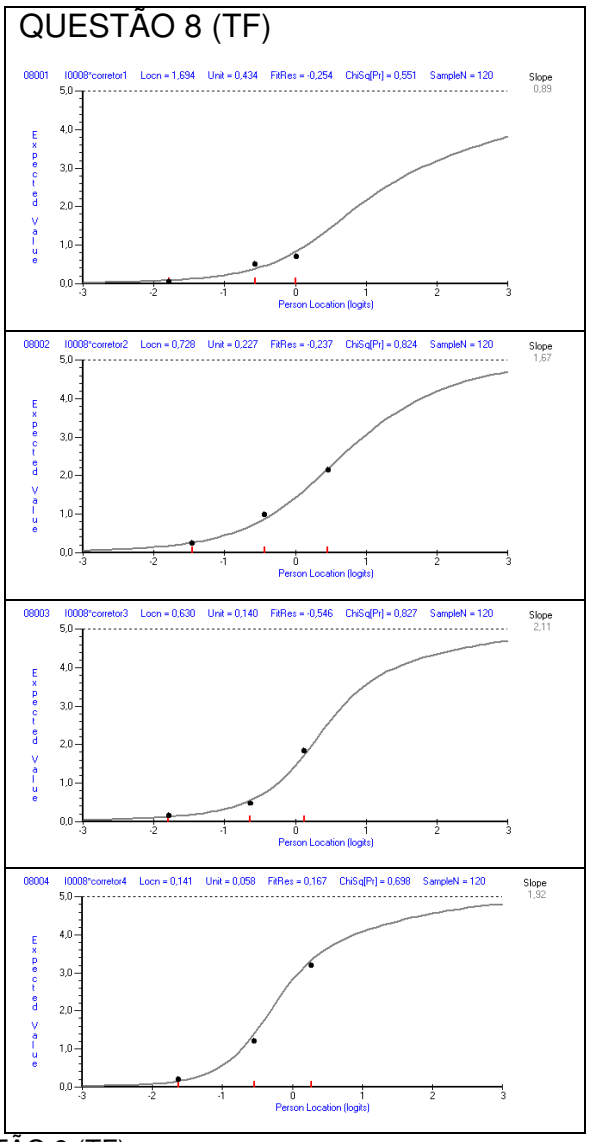


GRÁFICO 17 – QUESTÃO 8 (TF)
FONTE: AUTOR (2008)

A questão 8, que apresentou problemas de elaboração consideráveis no piloto, comprometendo significativamente a validade do teste e foi substituída. Pode-se observar que esta é uma questão de elevado grau de dificuldade, uma vez que o escore do grupo A só está localizado acima de 3,0 para um dos corretores. Percebe-se que o cor1 foi bastante rigoroso e o cor4 foi o mais leniente de todos, porém, sem o ser em demasia, visto que a diferença entre ele e os outros não é grande. Nota-se também que, em nenhum caso, a diferença entre os escores observados e os previstos pelas curvas é significativa e, ainda, que há uma boa discriminação entre os grupos, podendo-se considerar que esta também foi uma substituição bem sucedida.

É possível perceber pelos gráficos, que os grupos que realizaram os testes, de maneira geral, são compostos de leitores cuja proficiência de leitura ainda está longe daquela proposta para o leitor ideal dessa pesquisa135. A causa desse distanciamento pode ser tanto a falta de conhecimento de língua quanto a falta de experiência como leitor, entre tantas outras possibilidades que poderiam ser objeto de outro estudo que permitisse uma investigação mais detalhada da influência dos diversos fatores que dificultam o processo de compreensão em cada um desses grupos. O fato relevante nesse momento é que, independentemente da causa, a habilidade de compreensão de leitura não é muito elevada, nem mesmo no grupo A, do qual se esperaria um nível alto de proficiência de leitura. Observa-se que em nenhuma das questões a habilidade desse grupo está localizada acima de 1, no eixo horizontal. O posicionamento dos três grupos no piloto 1B está sempre concentrado bem próximo ao 0, entre -1 e +1, e no TF os grupos estão distribuídos entre -2 e +1 em todas as questões, com o grupo A sempre mais próximo do 0 do que de +1. Lembrando que o ponto 0 indica que a habilidade que o indivíduo demonstra corresponde a 50% de chance de acerto em uma questão de dificuldade média, tem-se que o grupo mais proficiente, entre os avaliados, tem pouco mais do que essa porcentagem de chance de acerto em todas as questões elaboradas. Isso demonstra que a habilidade de compreensão de leitura desses indivíduos está aquém do que se espera de alunos de cursos de mestrado e doutorado. Sendo assim, como argumentado no item 2.2, não é possível adotar o leitor ideal como sendo o leitor real desses testes, pois o nível de proficiência desses dois leitores é consideravelmente diferente. Esse fato, no entanto, não invalida o processo de avaliação e as análises realizadas; demonstra que a habilidade de compreensão de leitura necessária para a realização de um teste de suficiência como esses que foram elaborados, cujo construto está adequado aos propósitos do teste, deveria ser maior para estar mais próxima do ideal.

A análise contrastiva dos gráficos do 1B e do TF e a comparação entre os gráficos dos quatro corretores do TF permitem afirmar que as substituições e reformulações das questões, bem como a elaboração dos descritores e o treinamento dos corretores foram muito bem sucedidos. As anomalias

[135] Ver discussão sobre leitor ideal e leitor real no capítulo 2, item 2.2.

presentes nos gráficos do 1B desapareceram no TF em função do treinamento dos corretores e uso dos descritores, e a re-elaboração das questões mais problemáticas alcançou o resultado esperado. Tudo isso proporcionou maior validade e confiabilidade ao Teste Final.

6 CONSIDERAÇÕES FINAIS

O presente estudo teve a intenção de demonstrar a viabilidade e necessidade de utilização de análises estatísticas no processo de elaboração e correção dos testes de suficiência de leitura em inglês que são aplicados em candidatos ao, ,ou em alunos de mestrado e doutorado. Conforme o objetivo geral estabelecido apresentado no capítulo 1, item 1.1.1, o foco principal foi a elaboração de um teste, sua correção e a análise estatística dos dados obtidos com a utilização do Modelo Rasch. Foram elaborados quatro testes (Pilotos 1A, 1B, 1C e 1D), não apenas um, como proposto. Esses testes serviram de piloto para que um deles pudesse ser selecionado para análise estatística, re-elaboração, reaplicação e nova análise.

Na seleção do teste foi observado se, após a aplicação, o teste ainda se mostrava adequado quanto aos objetivos institucionais e se os resultados obtidos pela análise estatística possibilitavam a identificação dos problemas apresentados e uma comparação entre esses resultados e os obtidos depois da re-elaboração do teste. Sendo assim, o teste selecionado (Piloto 1B) foi corrigido e a sua qualidade foi analisada conforme a teoria sobre leitura e avaliação apresentada no levantamento bibliográfico. Durante a correção e a análise foram identificados os problemas e proposta uma re-elaboração, visando corrigi-los. Depois disso, o Teste Final (teste 1B re-elaborado) foi reaplicado em outro grupo de indivíduos, que não havia realizado nenhum dos quatro testes anteriores. Uma nova análise com a utilização do Modelo Rasch foi, então, realizada no Teste Final.

A pesquisa demonstrou claramente que a utilização de um instrumento de análise de dados com bases estatísticas, como o Modelo Rasch, permite que a qualidade dos testes elaborados aumente de forma significativa. Com isso, os testes ganham em validade e confiabilidade, na medida em que as questões propostas que apresentam problemas podem ser reelaboradas ou substituídas, à medida que forem identificadas.

A identificação das questões que apresentavam problemas no teste 1B foi feita, primeiramente, durante a correção, apresentada no capítulo 4 – item

4.3, e confirmada, posteriormente, pela análise dos gráficos, apresentada no capítulo 5 – item 5.2.1. A correção ou substituição das questões mostrou ter sido apropriada, uma vez que os resultados posteriores foram considerados melhores, conforme se pode observar na análise que se encontra no item 5.2.2. As informações geradas nos gráficos mostraram resultados bastante consistentes, tanto em relação à elaboração das questões quanto à correção realizada.

Tem-se, então, que os gráficos gerados pelo aplicativo possibilitam a identificação de problemas não só das questões, mas também de correção, mostrando onde ocorrem divergências entre os corretores. Embora não seja possível eliminar todos os problemas de correção, porque não se pode eliminar totalmente a subjetividade de processos como esse, a possibilidade de identificação desses problemas minimiza a interferência da subjetividade na correção, pois confere maior objetividade com a unicidade dos critérios adotados pelos corretores e, dessa forma, contribui para o aumento da confiabilidade do teste.

O treinamento dos corretores, conforme apontado pelos estudiosos citados no levantamento bibliográfico, teve sua relevância confirmada e mostrou que é perfeitamente possível que corretores diferentes consigam resultados semelhantes de correção com a utilização de critérios em comum. Os testes piloto foram corrigidos sem o estabelecimento de critérios e sem o treinamento dos dois corretores participantes, enquanto o teste final teve critérios estabelecidos e corretores treinados. O fato de o número de corretores ter sido aumentado para quatro no teste final poderia ter prejudicado a confiabilidade do teste, em função da subjetividade e diferenças entre os corretores, caso não houvesse treinamento. Além disso, o cuidado de fazer com que cada teste fosse corrigido por dois corretores diferentes garantiu que todos eles fossem corrigidos por dois corretores diferentes, como aconteceu no piloto, e que as correções pudessem ser comparadas, para identificar possíveis divergências significativas. Como conseqüência dos bons resultados obtidos a partir do treinamento dos corretores, pode-se deduzir que em testes com um grande número de examinandos é possível a utilização de questões discursivas, desde que os corretores sejam treinados adequadamente e que os

critérios estejam estabelecidos de forma clara, sem prejuízo da confiabilidade do teste, que poderia ocorrer em função da necessidade de se contar com / se precisar de vários corretores e das diferenças de correção,..

Em se tratando dos critérios de correção, pode-se perceber que o estabelecimento de descritores conferiu maior validade ao teste, uma vez que foi definido o que deveria ser avaliado na correção a partir do construto definido para a elaboração do teste, garantindo que fossem analisadas nas respostas as variáveis que se pretendia medir. O estabelecimento de descritores possibilitou, também, que o elaborador do teste tivesse controle sobre a diferença de interpretação que os corretores poderiam fazer das respostas em termos de conteúdo e garantiu que todos eles verificassem a presença ou ausência de informações consideradas relevantes.

A utilização de questões discursivas em lugar de questões objetivas mostrou-se muito eficaz aos propósitos pretendidos. Possibilitou que se verificasse a compreensão de cada leitor, sem conduzir ou limitar o leitor às opções de resposta determinadas, entre as quais ele teria que escolher, podendo contar com o fator sorte e exigindo que ele demonstrasse a sua compreensão de leitura de forma mais independente. Dessa forma, pode-se afirmar que o uso de questões discursivas também contribuiu com a validade do teste, na medida em que possibilitou que se avaliasse a habilidade do leitor em uma situação similar à que ele deve usar nos estudos realizados em mestrado e doutorado, nas quais o leitor deve ser capaz de entender o que lê e utilizar as informações adequadamente na elaboração do seu próprio texto, quer seja uma dissertação ou uma tese. Outro ponto positivo na elaboração das questões foi o seu embasamento em questões gerais, geralmente presentes em textos acadêmicos. Foi possível perceber que as questões podem variar o grau de dificuldade, dependendo da maneira como o texto foi estruturado. Assim foi possível elaborar questões de diferentes níveis de dificuldade. Pasquali afirma que não existe um nível de dificuldade ideal fixo para os itens de um teste, mas que o nível depende da finalidade do teste.

> Se for desejado um teste para selecionar os melhores ou determinar se um patamar ‘x’ de conhecimento foi atingido (como nos testes de referência a critério, então os itens devem todos apresentar o nível de dificuldade do patamar que se quer como critério de seleção ou acima

> dele. [...] Se, entretanto, o interesse consiste em avaliar a magnitude diferencial dos traços nos sujeitos de uma população, como geralmente é o caso, então uma distribuição mais equilibrada dos itens em termos de dificuldade é requerida. (2003, p. 127)

Considerando nessa afirmação de Pasquali, os resultados obtidos nos testes e a análise realizada pelo Modelo Rasch, a autora acredita que o nível de dificuldade estabelecido para as questões do teste proposto foi adequado aos propósitos para os quais ele foi elaborado.

Com a realização de todo esse trabalho de pesquisa a autora espera ter chamado a atenção dos profissionais que trabalham com avaliação de leitura em inglês para a possibilidade de realização da elaboração e análise de testes de leitura com a utilização dos conhecimentos teóricos sobre avaliação de leitura, juntamente com o Modelo Rasch e mostrado que o aumento de qualidade deles é real e observável. Além disso, com a elaboração, re-elaboração e análise dos testes a autora esperava ter lançado algumas bases para uma possível futura equalização de testes e construção de uma escala de medição, que possam conferir maior validade e confiabilidade aos testes de suficiência de leitura em inglês. Sob o ponto de vista da autora as possibilidades apontadas pelos resultados dessa pesquisa bem como muitas outras questões aqui levantadas são merecedoras de mais estudos.

6.1 LIMITAÇÕES DO ESTUDO

Durante a pesquisa foram encontradas algumas limitações que, embora não tenham prejudicado o estudo, são dignas de nota.

Uma limitação foi a dificuldade de encontrar um grande número de sujeitos para realizar os testes. A idéia inicial era que se pudesse obter o maior número possível de sujeitos, para que a pesquisa ficasse o mais similar possível à situação real dos testes de suficiência. No entanto, para os testes-piloto só foi possível contar com 94 sujeitos e no teste final com 120. Esse número é bem inferior ao número de candidatos que normalmente realizam o teste de suficiência na UFPR. É importante ressaltar, porém, que o número

reduzido, em relação à situação real desse teste, não interferiu na validade ou na confiabilidade que ele apresentou, pois foi o suficiente para que se fizesse de forma adequada todas as análises previstas e se atingisse os objetivos traçados inicialmente.

Também não foi possível a elaboração de mais um teste, diferente do teste final, sua posterior análise e posterior equalização dos dois testes. Isso seria desejável, porém, essa impossibilidade deveu-se à limitação institucional do tempo de realização de pesquisas de mestrado. Da mesma maneira, o tempo do estudo não possibilitou o levantamento e a análise de todos os problemas existentes no teste, para que pudessem ser corrigidos e reavaliados. No entanto, os problemas maiores e mais relevantes foram identificados e corrigidos com sucesso, conforme se pode observar no quinto capítulo.

Finalmente, não foi possível analisar os dados gerados pelo aplicativo em sua totalidade e em detalhes. Isso exigiria um tempo de estudo bem maior em função da necessidade de entendimento de certos conceitos e cálculos estatísticos fora do domínio da autora no presente momento.

A autora considera que as limitações do estudo não o comprometeram de forma significativa. A pesquisa foi realizada conforme o previsto, em detrimento das limitações, que, se não existissem, possibilitaria o aprofundamento de todas as questões levantadas, garantindo a realização de uma pesquisa muito mais completa e abrangente. Em função das limitações apresentadas, foram anotadas algumas sugestões para dar continuidade a essa pesquisa e/ou para a realização de outras.

6.2 SUGESTÕES PARA FUTURAS PESQUISAS

Entre tantas possibilidades de pesquisas que precisam ser realizadas sobre avaliações de leitura em LE, algumas sugestões para pesquisas futuras a partir da elaboração e resultados deste trabalho são dadas a seguir.

Uma das sugestões de pesquisa diz respeito ao fato da dificuldade de leitura observada nas respostas poder ser um problema de leitura ou um problema de conhecimento de língua. No caso desta pesquisa, conforme apontado pela autora, foi considerado que, em primeiro lugar, esta é uma questão que merece um estudo mais aprofundado em relação aos testes de suficiência, mas que deve ser feita em outro momento, em um estudo à parte. Em segundo lugar, em razão dessa necessidade de maiores estudos a autora partiu do princípio de que, independente do problema ser uma questão de conhecimento de língua ou de habilidade de leitura, a capacidade do leitor deve ser avaliada nos testes de suficiência. Neles, o que está sendo avaliada é a existência ou ausência de habilidade; não se está analisando a causa de uma ou de outra. Sendo assim, fica aqui uma sugestão para futuras pesquisas, com vistas a identificar a causa das dificuldades de leitura entre esse grupo de leitores em particular.

Outra sugestão de pesquisa é a verificação de problemas relacionados com as respostas, tais como:

a) a objetividade ao redigir a resposta – ela deve ser, de acordo com os propósitos, mais sucinta ou mais estendida;
b) a inclusão ou omissão das informações relevantes e inclusão de informações desnecessárias – que tratamento dar à corretas e às incorretas, de maneira a garantir uma correção justa e o que pode ser feito a esse respeito em termos de ensino/aprendizado de língua;
c) a capacidade de argumentação na resposta – relacionada diretamente à questão da escrita, como possibilitar, através do ensino da língua estrangeira a conexão entre a habilidade de leitura e a redação de uma resposta coerente e coesa;

Em relação à utilização do Modelo Rasch, pode-se destacar a necessidade de se aprofundar os estudos dos recursos disponíveis que podem contribuir para a validade e confiabilidade dos testes de leitura.

Finalmente, uma das sugestões mais importantes, dados os objetivos desta pesquisa, é a necessidade de se fazer mais estudos com a elaboração

de testes discursivos, que sejam analisados com o Modelo Rasch. Isso possibilitará a equalização desses testes e permitirá a elaboração de uma escala de medição, para garantir que os testes de suficiência de inglês tenham o mesmo grau de exigência em diferentes testes elaborados e que o nível de proficiência dos diferentes grupos de indivíduos avaliados seja o mesmo, independentemente dos diferentes momentos em que cada um deles é avaliado.

REFERÊNCIAS BIBLIOGRÁFICAS

ALDERSON, J. Charles. **Assessing reading.** Cambridge: Cambridge University Press, 2002.

ALDERSON, J. Charles. **Language Test Construction and Evaluation.** Cambridge: CUP, 1996.

BACHMAN, L. F. **Fundamental Considerations in Language Testing.** Oxford: Oxford University Press, 1991.

BACHMAN, L. F.; PALMER, A. S. **Language testing in practice.** Oxford: Oxford University Press, 1997.

BOND, T. G., FOX, C. M. **Applying The Rasch Model: Fundamental Measurement in the Human Sciences**. New Jersey: LEA, 2001.

BOND, T. G., FOX, C. M. **Applying The Rasch Model: Fundamental Measurement in the Human Sciences**. 2nd ed. New Jersey: LEA, 2007.

BROWN, H. D. **Teaching by principles: An interactive approach to language pedagogy.** 2nd ed. New York: Longman, 2001.

CARRELL, P. L. A View of Written Text as Communicative Interaction: Implications for Reading in a Second Language. In: DEVINE, J.; CARRELL, P. L.; ESKEY, D. E. (Ed.) **Research in Reading in English as a Second Language.** Washington, D. C.:TESOL, 1987.

CEPE. Resolução N.º 62/03 – Estabelece normas gerais únicas para os cursos de pós-graduação stricto sensu (mestrado acadêmico, mestrado profissional e doutorado) da Universidade Federal do Paraná. Disponível em: <http://www.ufpr.br/soc/pdf/cepe/cepe6203.pdf>. Acesso em: 30/11/2006

CEPE. Resolução N.º 38/05 - Altera os artigos 34 e 38 da Resolução nº 62/03-CEPE que estabelece normas gerais únicas para os cursos de pós-graduação stricto sensu (mestrado acadêmico, mestrado profissional e doutorado) da Universidade Federal do Paraná.- Disponível em <http://www.ufpr.br/soc/pdf/cepe/cepe3805.pdf>. Acesso em 30/11/2006

COHEN, A D. **Assessing language ability in the classroom.** 2nd ed. Boston: Heinle & Heinle, 1994.
CONTRUTO; CRITÉRIO; TESTES. In: DAVIES, Alan et al. **Dictionary of language testing.** Cambridge: Cambridge University Press, 2002, p. 31; 37; 81.

COSTA VAL, M. G. **Redação e textualidade.** 2ª Ed. São Paulo: Martins Fontes, 1999.

CRONBACH, L.J., MEEL, P.E. Construct Validity in Psychological Tests. **Psychological Bulletin**, N.º 52, 281-302, 1955.

DEL SIEGLE. **Interpret Raw Scores** – Standardized Scores – Educational Research. Disponível em: <http://www.gifted.uconn.edu/siegle/research/Normal/Interpret%20Raw%20Scores.html>. Acesso em 15/11/2006

DUKE, N. K. Comprehension of What for What: Comprehension as a Nonunitary Construct. In: PARIS, S. G. **Children's Reading Comprehension and Assessment.** Mahwah: Lawrence Erlbaum Associates, 2004**.** (p. 93-104)

ESCALA INTERVALAR; ESCALA NOMINAL; ESCALA ORDINAL. In: DAVIES, Alan et al. **Dictionary of language testing.** Cambridge: Cambridge University Press, 2002, p 89; 128; 137.

FOUCAMBERT, J. **A Leitura em Questão.** Tradução: Bruno Charles Magne. Porto Alegre: Artmed, 1994.

GOLDMAN, S. R., BISANZ, G. L. **Toward a Functional Analysis of Scientific Genres: Implications for Understanding and Learning Processes**. in: OTERO, José, LEÓN, José A., GRAESSER, Arthur C., **The Psychology of Science Text Comprehension.** New Jersey: LEA, 2002.

KATO, M. A. **O aprendizado da leitura.** São Paulo: Martins Fontes, 2007.

KINTSCH, W., KINTSCH, E. **Comprehension.** In: PARIS, S. G. **Children's Reading Comprehension and Assessment.** Mahwah: Lawrence Erlbaum Associates, 2004. (p. 71-92)

KOCH, I. V., ELIAS, V. M. **Ler e compreender: os sentidos do texto.** 2. Ed. São Paulo: Contexto, 2006.

MARTIN, M.T., **Standardized Scores**. 2008. Disponível em: http://www.azsba.org/static/index.cfm?contentID=202. Acesso em: 13/10/08.

MCNAMARA, T. **Language Testing.** Oxford: Oxford University Press, 2000.

MCNAMARA, T. **Measuring Second Language Performance.** London: Longman, 1996.

NERY, R. M. **Questões sobre questões de leitura.** Campinas, SP: Alínea, 2003.

NUNAN, D. **Research Methods in Language Learning.** Cambridge: Cambridge University Press, 2003.

NUTTALL, C. **Teaching reading skills in a foreign language**. Oxford: Heinemann, 2000.

OTERO, J., LEÓN, J. A., GRAESSER, A. C., **The Psychology of Science Text Comprehension.** New Jersey: LEA, 2002.

PARÂMETROS CURRICULARES NACIONAIS : **Terceiro e quarto ciclos do ensino fundamental: língua portuguesa.** Secretaria de Educação Fundamental. . Brasília :MEC/SEF, 1998.

PASQUALI, L. **Psicometria: Teoria dos testes na Psicologia e na Educação**. 2ª Ed. Petrópolis: Vozes, 2003.

PISA – Programme for International Student Assessment - 2000 RELATÓRIO NACIONAL - Brasília, Dezembro de 2001 http://www.inep.gov.br/download/internacional/pisa/PISA2000.pdf

PREL – Pacific Resources for Education and Learning. **A Focus on Comprehension.** 2005. Disponível em: <www.prel.org/programs/rel/rel.asp.> Acesso em: 07/07/2006

SNOW, C. E. **Reading for understanding: toward an R&D program in reading comprehension.** Santa Monica: RAND Education, 2002.

SWALES, J. M. **Genre Analysis. English in academic and research settings.** Cambridge: CUP, 2005.

TUMOLO, C. H. S. **Assessment of Reading in English as a Foreign Language: Investigating the Defensibility of Test Items**. Tese de Doutorado do Programa de Pós-Graduação em Letras/Inglês e Literatura Correspondente, Universidade Federal de Santa Catarina. Florianópolis, 2005.

UNIVERSIDADE FEDERAL DO PARANÁ – UFPR- Regimento Geral da Universidade Federal do Paraná. Disponível em <http://www.ufpr.br/soc/pdf/Reg%20Geral%20da%20UFPR%20alterado.pdf>. Acesso em 30/11/2006

VALIDADE CORRENTE; VALIDADE PREDITIVA; VALIDADE. In: DAVIES, Alan et al. **Dictionary of language testing.** Cambridge: Cambridge University Press, 2002, p. 30; 149; 221.

WRIGHT, B.D. Sample-free Test Calibration and Person Measurement**. MESA Research Memorandum Number 1**. Disponível em <http://www.rash.org/memo1.htm>. Acesso em: 01/01/2002

WRIGHT, B.D., LINACRE, J.M. Observations are Always Ordinal; Measurements, however, Must be Interval**. MESA Research Memorandum Number 44. MESA PSYCHOMETRIC LABORATORY**, 1989. “This appeared in *Archives of Physical Medicine and Rehabilitation* 70 (12) PP. 857-860, November 1989” Disponível em <http://www.rasch.org/memo44.htm>.

WRIGHT, B.D., STONE, M.H., **The measurement model. *Best Test Design: Rasch Measurement***, Mesa Press: Chicago, **1979** (p. 1-17). In: Modern Test Theory in Psychology and Education I – Reader. Perth, Western Australia: Murdoch University, 2002.

GLOSSÁRIO

ANÁLISE POLITÔMICA DE DADOS: significa que para cada item analisado estarão envolvidas diferentes categorias ordenadas de resposta, com pontuação diferente para cada categoria.

AVALIAÇÃO: Avaliação pode ser definida como a coleta sistemática de informações para o propósito de tomada de decisões.

"Evaluation can be defined as the systematic gathering of information for the purpose of making decisions."

CONFIABILIDADE: Confiabilidade é uma qualidade dos *escores* do teste e, conseqüentemente, tem a ver com a consistência da medição em diferentes tempos, formas de teste, corretores, e outras características do contexto de medição.

CONSTRUTO: o traço ou traços que um teste pretende medir. Um construto pode ser definido como uma capacidade ou um conjunto de capacidades que estarão refletidas no desempenho de um teste, e sobre as quais poderão ser feitas inferências com base na pontuação do teste. Um construto é geralmente definido em termos de uma teoria; no caso da linguagem, uma teoria da linguagem. Um teste, então, representa uma operacionalização da teoria. Validação do construto envolve uma investigação do que um teste na verdade mede e tenta explicar o construto.

CRITÉRIO: (1) desempenho na situação fora do teste; (2) um aspecto no qual o desempenho é julgado, por exemplo, *fluência, adequação*.

DESCRITOR: diz-se de ou parâmetro de registro de dados que possibilita que todos os registros referentes a um mesmo assunto sejam identificados. (Houaiss-parcial)

EQUALIZAÇÃO: O processo de estabelecer a equivalência de itens ou tarefas individuais de testes ou de testes completos. Quando vai ser feita uma comparação entre dois escores fornecidos por dois ou mais testes ou formas de teste, um procedimento é normalmente usado, por meio do qual ou os testes são dados ao mesmo grupo e examinandos, ou um número de amostras de examinandos, com características similares cada, fazem um teste diferente. Um ponto comum de referência é então estabelecido, e uma escala padronizada comum pode ser construída.

"The process of establishing the equivalence of individual test items or tasks or of entire tests.When a comparison is to be made between scores provided by two or more tests or test forms, a procedure is normally used whereby either the tests are given to the same group of test takers, or a number of samples of test takers with similar characteristics each takes a different test. A common point of reference is thus established, and a common standardized scale may be constructed.An alternative approach is to establish the characteristics (difficulty, discriminability, reliability) of individual test items or tasks, using, for example, IRT analysis, and thus to develop an item bank. Once a large enough pool, or bank, is developed, multiple forms of a test of known difficulty may then be constructed from these items. Items may be selected in order to produce tests of equal difficulty (alternate forms – horizontal equating); or of increasing difficulty (where test A is for beginners, for example, test B for intermediate learners, and test C for advanced learners – vertical equating)." (Dictionary of Language Test – p. 53)

EQUALIZAÇÃO DE TESTE: O processo de comparar a dificuldade de duas ou mais formas de um teste, para estabelecer sua equivalência. Isso é importante quando formas paralelas de um teste padronizado serão administradas em diferentes ocasiões.

"The process of comparing the difficulty of two or more forms of a test, in order to establish their equivalence. This is important when parallel forms of a standardized test are to be administered on different occasions." (Dictionary of Language Test – p. 198)

EQUIVALÊNCIA DE TESTE: A relação entre duas ou mais formas do mesmo teste. As formas de testes podem ser equalizadas ou equivalentes.

ERROS DE AMOSTRAGEM: São erros que ocorrem em virtude da pesquisa empírica geralmente não poder ser realizada com todos os membros de uma população, eventos ou objetos.

ERROS DE OBSERVAÇÃO: (1) erros instrumentais devidos a inadequações do instrumento de observação, (2) erros pessoais devidos às diferentes maneiras de cada pessoa reagir, (3) erros sistemáticos devidos a algum fator sistemático não controlado, como por exemplo, medir a temperatura em nível diferente da do mar, e (4) erros aleatórios, que não têm causa conhecida ou cognoscível.

ESCALA: Um sistema graduado de níveis. A teoria da mensuração faz uso de quatro tipos de escala: Nominal; ordinal; intervalar; proporcional/de proporção.

ESCALA INTERVALAR: Uma escala intervalar é uma numeração de diferentes níveis em que as distâncias, ou intervalos, entre os níveis são iguais. Isto é, além da ordenação que caracteriza escalas ordinais, escalas intervalares consistem de distâncias ou intervalos iguais entre os níveis ordenados. Escalas intervalares, portanto, possuem as propriedades da distinção, ordenação e intervalos iguais.

ESCALA NOMINAL: Aquela que consiste na contagem de ocorrências de atributos mutuamente exclusivos. Portanto, é mais a medida da freqüência da ocorrência de um atributo do que o quanto dele está presente.

ESCALA ORDINAL: Uma escala que ordena objetos de acordo com sua relação uns com os outros. Os pontos na escala ficam em uma relação de ‘mais que’ ou ‘menos que’ entre si. Enquanto uma escala ordinal é capaz de ordenar itens em relação uns aos outros, o tamanho do aumento entre dois pontos adjacentes não pode ser presumido como sendo o mesmo. Uma escala ordinal, conseqüentemente, não pode informar sobre o grau de diferença entre dois itens, por exemplo, a diferença de habilidade entre dois candidatos.

ESCORE BRUTO: escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste.

ESCORE PADRONIZADO: Uma transformação de escores brutos que fornece uma medida com uma posição relativa em um grupo e permite a comparação entre os escores brutos de diferentes distribuições, por exemplo, de testes de extensões diferentes. Isso é feito pela conversão de um escore bruto em uma organização padrão de referência que é expressa em termos da sua posição relativa na distribuição dos escores. O escore z é o escore padronizado mais comumente usado. (Dictionary of Language Test – p. 186-7)

ESCORE VERDADEIRO: escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio escore bruto se não houvesse o erro de medida. Formas equivalentes de testes são elaboradas a partir das mesmas especificações de teste, para medir as mesmas habilidades. Espera-se que os escores nas duas formas de teste sejam equivalentes; que a média e a variância sejam iguais. Em testes equalizados, por outro lado, o objetivo não é apresentar versões paralelas de um teste com distribuição equivalente de escores, mas transformar os escores de ambos os testes em uma escala em comum, que permita comparações entre os testes.
"The relationship between two or more forms of the same test. Test forms may equated or equivalent.Equivalent forms of tests are constructed from the same test specifications in order to measure the same skills. Scores on the two test forms will be expected to be equivalent; the mean and variance will be equal.In equated tests, on the other hand, the aim is not to produce parallel versions of a test with equivalent score distribution, but to transform scores from both tests onto a common scale which allows for comparison across the tests." (Dictionary of Language Test – p. 198-9)

FUNÇÃO CARACTERÍSTICA DO ITEM OU CURVA CARACTERÍSTICA DO ITEM (CCI): equação monotônica crescente que descreve a relação entre o desempenho na tarefa em questão e o conjunto de traços latentes.

LEITURA INTERATIVA: uso complementar das abordagens top-down e bottom-up.

MENSURAÇÃO: Mensuração nas ciências sociais é o processo de quantificar as características de pessoas de acordo com procedimentos e regras

explícitas. Esta definição inclui três aspectos distintos: quantificação, características, e procedimentos e regras explícitas.

"Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules. This definition includes three distinguishing features: quantification, characteristics, and explicit rules and procedures."

MODELO MENTAL: *(mental model)* é a representação mental interna que um indivíduo faz de uma realidade externa, para poder entendê-la. É uma representação que assume forma análoga à da informação que lhe deu origem, embora seja incompleta, não represente com exatidão a realidade e esteja sujeita a mudanças.

MODELO SITUACIONAL: (*situation model*) é uma representação da situação descrita no texto; isto é, uma representação mental do texto, que o leitor constrói, e que requer uma integração entre a informação do texto, o conhecimento prévio do leitor e os seus objetivos de leitura.

PARÂMETRO: 5- MAT característica diferencial que é passível de mensuramento ou direta ou indiretamente. 10- GRAM cada um dos valores admitidos por um princípio lingüístico na sua aplicação cada língua particular. 11- PSIC qualquer constante qu4e define a curva da equação de determinado evento psicológico. 12- PSIC curva que representa um desempenho sob condições experimentais específicas. (Houaiss).

PROCESSAMENTO ASCENDENTE: O processamento ascendente (*bottom-up*) faz uso linear e indutivo das informações visuais, lingüísticas, e sua abordagem é composicional, isto é, constrói o significado através da análise e síntese do significado das partes.

PROCESSAMENTO DESCENDENTE: O processamento descendente (*top-down*) é uma abordagem não-linear, que faz uso intensivo e dedutivo de informações não-visuais e cuja direção é da macro para a microestrutura e da função para a forma.

PSICOMETRIA: teoria de medida utilizada Na medição quantitativa de aspectos do comportamento humano, ou seja, de construtos psicológicos, assim como de qualquer traço da capacidade intelectual do ser humano.

TEORIA CLÁSSICA DOS TESTES (TCT): A Teoria Clássica dos Testes (TCT) se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total (T).

TEORIA DE RESPOSTA AO ITEM (TRI): A Teoria de resposta não se interessa pelo escore total do teste, mas por cada item de que ele se compõe. Além disso, essa teoria procura saber a probabilidade de acerto e erro de cada item e quais os fatores que afetam essa probabilidade.

TESTE: um teste é um instrumento de medição elaborado para obter uma amostra do comportamento de um indivíduo. Como um tipo de medição, um teste necessariamente quantifica características de indivíduos de acordo com procedimentos explícitos.

"a test is a measurement instrument designed to elicit a specific sample of an individual's behavior. As one type of measurement, a test necessarily quantifies characteristics of individuals according to explicit procedures."

TRAÇO OU TRAÇO LATENTE: Traços ou traços latentes são usados nesta pesquisa com o significado de: uma variável não observável; um processo psicológico; atributos que não estão sujeitos à observação empírica e que, por isso, precisa ser observado por meio dos comportamentos que são a eles atribuídos, para que possam ser 'objetos' de uma abordagem científica.

UNIDIMENSIONALIDADE: a possibilidade de se medir apenas um atributo separadamente, sem que ele dependa ou sofra a influência de algum outro.

VALIDAÇÃO: O processo de estabelecer a validade de um teste, que é uma das preocupações básicas de avaliação de linguagem. (Dictionary of Language Test – p. 220)

VALIDADE: é a qualidade que mais afeta o valor de um teste, anterior a, embora dependente da **confiabilidade**. Uma medida é válida se faz o que ela tem a intenção de fazer, que é tipicamente agir como um indicador de um conceito abstrato (por exemplo, altura, peso, tempo, etc.) que ela afirma medir. A validade de um teste de linguagem, portanto, é estabelecida pelo grau com o qual ele é bem sucedido em provar uma representação concreta precisa de um conceito abstrato (por exemplo, **proficiência**, **realização**, **aptidão**).

APÊNDICES

1– TESTES ELABORADOS A PARTIR DE RESUMOS DE TESES

Teste 1

READING COMPREHENSION I

1ST SEMESTER - 2007

Thesis (M.Ed.)--University of Windsor, Canada, 2005

By Maureen Harris

STUDENT ABSTRACT: While a growing body of research reveals the beneficial effects of music on education performance the value of music in educating the young child is not being recognized, particularly in the area of Montessori education. This study was an experimental design using a two-group post-test comparison. A sample of 200 Montessori students aged 3 to 5-years-old were selected and randomly placed in one of two groups. The experimental treatment was an "in-house" music enriched Montessori program and children participated in 3 half-hour sessions weekly, for 6 months. This program was designed from appropriate early childhood educational pedagogies and was sequenced in order to teach concepts of pitch, dynamics, duration, timbre, and form. The instrument used to measure mathematical achievement was the Test of Early Mathematics Ability-3 to determine if the independent variable, music instruction had any effect on students' mathematics test scores, the dependent variable. The results showed that subjects who received music enriched Montessori instruction had significantly higher mathematics scores. When compared by age group, 3 year-old students had higher scores than either the 4 or 5 year-old children.

Responda as questões abaixo em português. De acordo com o texto:

1) Qual é o problema de pesquisa proposto pela pesquisadora?
2) Qual foi o objetivo da pesquisa?
3) Que conclusão a pesquisadora pode tirar a partir do estudo realizado?

Teste 2

READING COMPREHENSION I

1ST SEMESTER - 2007

Thesis (Ph.D.)--Simon Fraser University, Canada , 2003

By Heather Mohan Van Heerden

STUDENT ABSTRACT: This thesis explores connections between the arts, human suffering, and healing through an in-depth examination of my personal journey as a music therapist in palliative care. In particular, it explores what I have learned about grief and healing through my clinical work with children who have lost a loved one to a terminal illness. This inquiry led me towards the articulation of a personal model of music therapy practice, rooted in a holistic view of healing, which has emerged from my clinical experiences in palliative care, and from a phenomenological analysis of those experiences. The guiding premise behind this inquiry was not a search for 'truths' or causal links, but rather a search for a greater depth of understanding of the meaning of lived experiences of grief and healing in arts therapy from both the participants' perspectives and from the therapist's perspective. Meaning was explored in five thematic dimensions, each of which is examined in more detail in the five chapters that make up the body of the work: dissonance, wounding, healing, grieving, and love. In addition, dialogues with five bereaved children who participated in an arts therapy support group were conducted by the clinician-researcher in order to gain more insight into their experiences of grief and healing. It was discovered that the group provided a community of belonging for grieving children; a space where they felt safe, comforted, relaxed and understood. Arts therapists have struggled in Canada for recognition of their work, and for ways to convey its meaning and depth through research to other health care professionals. In addition, there has been very little written in the academic literature about the use of music or arts therapy with grieving children. One of our greatest challenges is the translation of our experiences in music, movement, and visual imagery into words. The arts happen in the 'in-between

spaces'---those spaces which are impossible to measure and difficult to capture in discursive language. Meanings in this realm are conveyed intuitively rather than propositionally. In an attempt to shift from traditional positivistic paradigms of measurement, this thesis engages a phenomenological approach of inquiry, incorporating the arts as an integral part of the research process. Narratives, stories, images and poetic forms which illuminate connections between the arts and healing are used in order to invite the reader inside the world of lived experience, and to illustrate a particular way of understanding the arts in therapy and in education which promotes a holistic conception of healing.

Responda as questões abaixo em português. De acordo com o texto:

1) Qual o objeto de estudo desta tese?
2) O que o estudo realizado pretendia entender?
3) Qual a posição dos terapeutas ”de artes” no Canadá?
4) Explique qual é a dificuldade dos terapeutas mencionada no texto e porque ela existe.

Teste 3

READING

COMPREHENSION I

Thesis (M.C.P.)--University of Manitoba, Canada, 2006.

By Khawja A Latif

STUDENT ABSTRACT: In the wake of diminishing support for social housing and an ever increasing demand for suburban living with improved transit services, inner City neighborhoods have been declining in Winnipeg. In an effort to revitalize inner city neighborhoods and create demand for housing, the City of Winnipeg and the Province of Manitoba are investing in the rehabilitation of inner city neighborhoods. West Broadway is one such neighborhoods, where location creates a higher demand for new smaller households and existing types of housing can supply a part of the accommodation in the future through construction, alterations and renovations of old houses. Recently, provision has been created for the use of Tax Increment Financing (TIF) for community development financing, which can be made available for financially unfeasible renovation, alteration, maintenance and reconstruction for the purpose of revitalization of the area. The practicum explores the feasibility of establishing a TIF for West Broadway. It investigates the financial feasibility of return of the renovation investments in the form of property taxes to the City and to the Province from West Broadway neighborhood. The study reveals that the returns in the form of incremental education and municipal taxes from 2000 to 2009 are not feasible against investments made from 2000 to 2005; however, a TIF project life of at least fifteen years is feasible. The financial feasibility, albeit depends on certain conditions. Based on the study, some conclusions were drawn and recommendations for the adoption of a TIF for the West Broadway are made for instituting a self financed TIF district for the West Broadway neighborhood.

Responda as questões abaixo em português. De acordo com o texto:

1) Qual foi o problema que motivou esta pesquisa?
2) a) Que tipo de provisão foi, recentemente, criada?
   b) E para quê ela está disponível?
3) Qual o objetivo da pesquisa?
4) O que a pesquisa revelou?

Teste 4

READING

COMPREHENSION I

Thesis (M.A.)--University of New Brunswick, Canada, 2004

By Brian Christopher Gallant

STUDENT ABSTRACT: The economic crisis of the 1930s required the federal government, under the leadership of R. B. Bennett, to devise strategies to deal with unemployment. One such measure the government devised was the creation of unemployment relief camps, where single unemployed men worked in return for food, clothing, shelter, and an allowance of 20 cents per day. Seven of these relief projects were established in New Brunswick under the direction of the Department of National Defense. In western Canada, these camps came to be seen as hotbeds of unrest, but the New Brunswick camps do not seem to fit this vision of camp life. Despite the negative reputation of the camps in the historiography, they provided much-needed relief to New Brunswick men and also offered benefits including opportunities for free education and health care. There were no major disturbances in the New Brunswick camps, and some men who worked in these camps remember them fondly. This study contributes to the historiography of the unemployment relief camps and of the Great Depression in the Maritime Provinces, breaking away from the negative stereotypes surrounding this government programme.

Na 1ª metade do séc. XX o governo Canadense criou algumas "instituições". Responda as questões abaixo em português:

1) Que "instituições" eram essas?
2) A quê se destinavam?
3) Por que foram criadas?
4) Qual o funcionamento básico citado no texto?
5) Qual era a opinião sobre essas "instituições" no oeste do país?
6) A conclusão do pesquisador está de acordo com a opinião mencionada na questão acima? Justifique.

2– TESTES-PILOTO APLICADOS

PILOTO 1A

Sufficiency Test – Pilot 1A

Student: _____

Institution: _____ **Course:** _____

Teacher: Graziella Lapkoski - UFPR **Grade:**_____ **Nov. 07**

***Permitido o uso do dicionário!**

****Teste deve ser respondido em português!** ***Usar caneta azul ou preta!***

Tempo de duração previsto: entre 1:30h e 2:00h.

renatal thumb sucking is related to postnatal handedness

Peter G. Hepper , Deborah L. Wells and Catherine Lynch

School of Psychology, The Queen's University, Belfast, BT7 1NN, UK

Received 5 July 2004; revised 24 August 2004; accepted 27 August 2004.

Available online 17 November 2004. doi:10.1016/j.neuropsychologia.2004.08.009

Neuropsychologia

Volume 43, Issue 3, 2005, Pages 313-315

Abstract

This study followed-up 75 individuals who were observed sucking their thumb as fetuses and examined their handedness, assessed by a modified version of the Edinburgh Handedness Inventory, at 10–12 years of age. Of 60 right-handed fetuses, all were right-handed postnatally; 10 of 15 left-handed fetuses were left-handed and five right-handed. Male left thumb sucking fetuses were more likely to be right-handed children than females. The study indicates that the prenatal exhibition of lateralised motor behaviour, in this case thumb sucking, is indeed related to postnatal handedness, perhaps more strongly for right 'handed' fetuses than left 'handed' fetuses.

Keywords: Handedness; Laterality; Fetus; Child; Motor behaviour

Article Outline

1. Introduction

2. Method

2.1. Participants

2.2. Procedure

2.3. Analysis and results

3. Discussion

Acknowledgements

References

1. Introduction

Observations of handedness, or lateralised motor behaviour, have largely been restricted to the postnatal period (McManus, 2002). Recent advances in ultrasound technology have enabled the motor behaviour of the fetus to be observed, in real-time, from its earliest appearance at 8 weeks of gestation (Nijhuis, 1992). A number of reports have examined whether the fetus exhibits laterality in its motor behaviour. An early report (Hepper, Shahidullah, & White, 1991) examining fetal thumb sucking reported that approximately 90% of fetuses sucked their right thumb and 10% their left. This bias was observed at 15 weeks of gestation, the earliest age studied. A study involving arm movements observed fetuses moved their right arms more than their left at 10 weeks of gestation (Hepper, McCartney, & Shannon, 1998), the earliest age at which isolated arm movements can be observed. This finding has recently been confirmed using four dimensional ultrasound (Kurjak et al., 2002). A study of fetal head position observed that at 38 weeks gestation fetuses were more likely to have their head turned to the right than the left (Ververs, de Vries, van Geijn, & Hopkins, 1994). A lateralised bias was not found in hand–face contact in fetuses (de Vries et al., 2001) although only 10 fetuses were observed in this study. Most recently, fMEG studies have been employed on the fetus in the third trimester and it has been observed that, using auditory evoked responses, there are functional hemispheric asymmetries in auditory evoked activity in the fetal cortex (Schleussner et al., 2004).

There has been much speculation on the interaction between handedness and hemispheric specialisation (Annett, 1985; Geshwind & Galburda, 1985; Porac & Cohen, 1981; Previc, 1991). The studies reported above indicate the presence of 'handedness' early in gestation and also functional hemispheric specialisation in later gestation. Questions remain however, as to whether the prenatal exhibition of handedness has any relationship to handedness present after birth. There is some indication that fetuses maintain their preference for both thumb sucked (Hepper et al., 1991) and arm moved (McCartney & Hepper, 1999) during gestation suggesting this prenatal behaviour displays a consistent lateralisation. In the study of fetal thumb sucking (Hepper et al., 1991), it was found that fetuses who sucked their right thumb were more likely to turn their head to the right and those who sucked their left thumb to turn their heads to the left when examined 2–4 days after birth suggesting prenatal handedness may relate to postnatal lateralised motor behaviour. It is the aim of this study to examine the handedness of children at 10–12 years of age who had previously been observed as fetuses sucking their left or right thumb.

2. Method

2.1. Participants

Eighty children (and their parents) who had previously been observed as fetuses (Hepper et al., 1991) were contacted to see if they would be willing to participate in a follow-up study and complete a questionnaire to assess their handedness. Seventy-five agreed and their data is reported here. Of those not agreeing, two were in the process of moving house and three were undertaking exams, and declined. Four of these were right-handed and one

left-handed. Of the seventy-five children who participated in the study, 60 sucked their right thumb as fetuses and 15 their left. Children were aged between 10 and 12 years of age at their participation in this study.

2.2. Procedure

Handedness was assessed in the children by means of a modified Edinburgh Handedness Inventory (Oldfield, 1971). Ten questions were asked and subjects had to indicate whether they used their left or right hand for the task. The tasks were: writing; drawing; throwing a ball; holding scissors; using a TV or other remote control; holding a toothbrush; holding a knife without a fork; holding a spoon; holding a can or bottle to open it; writing text messages on a mobile phone. Subsequent analysis revealed that only 49 of the children used a mobile phone and hence this question was dropped from the analysis. For each subject, the number of times the individual responded with a left hand response was recorded. If all of the responses were right hand, the child scored 0 and if all the responses were left hand, the child scored 9.

2.3. Analysis and results

Children were deemed right-handed if they scored 4 or fewer responses and left-handed if they scored 5 or more. Of the 60 children who sucked their right thumb as fetuses, all ($n = 60$) were classified as right-handed at 10–12 years of age. Of 15 children who sucked their left thumb as fetuses, five were right-handed and 10 were left-handed at 10–12 years of age. The results clearly indicate right thumb sucking prenatally relates to subsequent right-handedness (100%) whereas the association between left thumb sucking and subsequent left-handedness is present but not as strong (66%). An examination of the risk ratio (as one cell = 0 an odds ratio, or tests based on this, would be inappropriate to undertake) reveals right thumb suckers are three times more likely to be right-handed than left thumb suckers to be left-handed ($p < 0.0001$).

Children were further subdivided into strongly or weakly right- and left-handed. An individual was considered to be strongly left-handed if they scored 9 and weakly left-handed if they scored 5–8. A child was considered strongly right-handed if they scored 0 and weakly right-handed if they scored 1–4. The results are presented in Table 1. A Fisher exact test examining the prenatal thumb sucked and the number of individuals exhibiting a strong or weak preference for the same hand at 10–12 years revealed a significant difference ($p < 0.001$ whether excluding the five left thumb suckers who were right-handed as children ($n = 70$) or including them as not strongly left ($n = 75$)). Fetuses who sucked their right thumb who significantly more likely to exhibit a strong preference for the same hand as children than were fetuses who sucked their left thumb.

Table 1.
The number of individuals who were classified as either strongly or weakly left and right-handed according to the thumb sucked prenatally

		Handedness at 10–12 years			
		Strongly left	Weakly left	Weakly right	Strongly right
Thumb sucked	Right	0	0	19	41
prenatally	Left	0	10	5	0

Table 2 presents the numbers of individuals scoring nine to zero left hand responses, broken down by thumb sucked as a fetus and sex. Observation of this table indicates that there were no sex differences amongst right-handed fetal thumb suckers and subsequent handedness but a slight difference in left-handed fetal thumb suckers. Here, more males moved from being left-handed as fetuses to right-handed as children compared to females. Numbers are, however, small and the observation requires very cautious interpretation and further study.

Table 2.

The number of individuals who scored nine to zero left hand responses according to their sex and the thumb sucked as a fetus

Thumb sucked		No. of left hand responses									
		9	8	7	6	5	4	3	2	1	0
R	Male						2	1	2	5	22
R	Female						1	2	1	5	19
R	Total						3	3	3	10	41
L	Male		2	2			1	1	1	1	
L	Female		2	2	2		1				
L	Total		4	4	2		2	1	1	1	

3. Discussion

The results indicate that prenatal lateralised motor behaviour is predictive of postnatal lateralised motor behaviour, i.e., prenatal handedness (as observed by thumb sucking) is strongly related to postnatal handedness measured by a modified Edinburgh Handedness Inventory. The results appear stronger for prenatal observations of right-handedness than left-handedness. Table 3 presents, for each item in the inventory, the percentage

number of individuals who use the hand opposite to that of the thumb sucked as a fetus. There is some suggestion that certain items may be less appropriate for assessing handedness reflecting a cultural bias, e.g., use of scissors. This may contribute to the poorer prediction of handedness for left-handers observed here. There is a suggestion in the results that the patterns of behaviour may be more continuous in females than males, but numbers here are small and a greater sample is needed to confirm this. Overall, these results suggest that there is continuity between the handedness observed before birth and that exhibited after birth.

Table 3.
The percentage no. of individuals who scored the opposite response to their thumb sucked prenatally for each of the items in the handedness inventory for those who sucked their right and left thumb prenatally

Item	Right thumb fetus	Left thumb fetus
Write	1.7	53.3
Draw	1.7	40
Throw	16.7	33
Scissors	0	60
TV remote	1.7	26.7
Toothbrush	1.7	26.7
Knife	8.3	26.7
Spoon	11.7	33.3
Can open	18.3	26.7

The results of this study suggest that the previous observations of lateralised motor behaviour (e.g., Hepper et al., 1998 and Ververs et al., 1994) in the prenatal period are indeed early (the earliest possible) manifestations of handedness prevalent and familiar after birth. These observations accompanied by the report of a functional cortical asymmetry in later pregnancy (Schleussner et al., 2004) indicate that the developmental ontogenesis of lateralised motor behaviour, and possibly functional cerebral asymmetries, occurs very early in development indeed. The underlying causation of handedness so early in gestation remains elusive but it is likely that both genetic and environmental factors play a role. Whatever the underlying mediation, handedness, a fundamental feature of behaviour after birth, appears to be a fundamental feature of behaviour before birth. The interaction between handedness and cerebral functional and physical asymmetry has yet to be resolved but it would appear this interaction is present from early in gestation. Further examination of prenatal laterality is warranted and may throw light on the proximate and ultimate causation of laterality.

Acknowledgements

The support of the School of Psychology, Queen's University and Royal Maternity Hospital, Belfast is gratefully acknowledged. I am grateful for the statistical advice of Martin Dempster and the comments of the reviewers.

Responda as questões abaixo em português:

1) Qual o objetivo da pesquisa?(1,0)

2) Que metodologia os pesquisadores utilizaram? (2,0)

3) Numa segunda fase da pesquisa, foram feitas dez perguntas aos sujeitos. (2,0)
   a) Qual o objetivo das perguntas?

   b) Por que uma delas foi desconsiderada?

4) Quais foram os resultados obtidos? (1,0)

5) Na pesquisa, o sexo dos bebês também foi considerado. O que os pesquisadores observaram?(1,0)

6) A pesquisa sugere que nem todos os itens mostraram-se adequados na indicação de “handedness”. (2,0)
   a) Por que?

   b) Qual o exemplo citado?

7) Esses resultados obtidos foram considerados conclusivos? _____ Por que? (1,0)

PILOTO 1B (e TESTE FINAL)

Sufficiency Test – Pilot 1B

Student: _____

Institution: _____ **Course:** _____

Teacher: Graziella Lapkoski - UFPR **Grade:** _____ **Nov. 07**

***Permitido o uso do dicionário!**

****Teste deve ser respondido em português!** ***Usar caneta azul ou preta!***

Tempo de duração previsto: entre 1:30h e 2:00h.

Insect cells for human food

M.C. Verkerk[a, b], J. Tramper[a], J.C.M. van Trijp[b] and D.E. Martens[a]

[a]Department of Agrotechnology and Food Sciences, Food and Bioprocess Engineering Group, Wageningen University, Bomenweg 2, 6703 HD, Wageningen, The Netherlands

[b]Department of Social Sciences, Marketing and Consumer Behaviour, Wageningen University, Hollandseweg 1, 6706 KN, Wageningen, The Netherlands

Available online 23 November 2006.

Biotechnology Advances

Volume 25, Issue 2, March-April 2007, Pages 198-202

Abstract

There is a need for novel protein sources. Insects are a possible interesting source of protein. They are nutritious in terms of protein (40–75 g/100g dry weight) and minerals. Insect protein is of high quality and has a high digestibility (77–98%) and concentration of essential amino acids (46–96% of the nutritional profile). Also insect cells may be a promising novel source of protein. Choice of cell line, growth conditions and use of the baculovirus expression system opens up possibilities to engineer the nutritional value of the biomass. The technological limits as well as consumer acceptance of insect cell based food remains to be investigated.

Keywords: Insects; Insect cell culture; Novel proteins; Consumer behaviour

Article Outline

1. Introduction

2. Novel proteins on the market: failure or success?

3. Insects as food

4. Nutritional value of insects

5. Insect cells from bioreactors

6. Consumer point of view

7. Concluding remarks

Acknowledgements

References

1. Introduction

In the light of population growth and the increase in welfare worldwide, there is a need for novel protein sources as an alternative for meat production. In general protein sources should be safe, nutritious, flexible, reliable as well as consumer accepted. These aspects are even more important for novel protein sources. Predictions are that in the next decades to come around 40% of traditional meat consumption will be replaced by novel protein sources (Kuijer and Wielenga, 1999). In this article the potential of insect biomass is discussed.

2. Novel proteins on the market: failure or success?

The production of novel protein produced by single cells was brought up by the oil companies rather than the food companies in the 1950s. They used waste/by-products as a substrate for the single cells to produce protein.

British Petroleum brought Tropina on the market, which is protein derived from yeast growing on alkanes. For 12 years Tropina was tested for toxicity and carcinogenicity and no adverse effects could be demonstrated. Nevertheless, the consumer questioned the safe status of Tropina. People were afraid that the product contained aromatic hydrocarbons. Opposition in Japan even led to a total ban of petrochemical derived proteins. In Europe governments required further research. These studies showed that the product was not carcinogenic. Based on these findings the use of Tropina was allowed in limited amounts but still only for export. However, when the price of oil increased the substrate needed for petrochemical protein became relatively expensive. Consequently, BP decided to stop the production of Tropina because it could not compete with the price of soy (Israelidis, 1988).

ICI had to cope with a similar problem. They introduced Pruteen, a protein produced by bacteria with methanol as a substrate. Pruteen had a crude protein content of 72% and was marketed as a highly balanced protein source for feed. ICI had commissioned a 60,000 t/year plant to produce Pruteen. Again, when the price of oil increased Pruteen was not able to compete with alternative protein sources, such as soy and fish. Despite the successful engineering of Pruteen it was not economic to produce (Israelidis, 1988).

A more successful novel protein is mycoprotein. The mycoprotein in the products of Quorn is produced by a Fusarium that grows on molasses (Israelidis, 1988). The proteins from this fungus are allowed for human consumption. In 1986 Quorn entered the British market, selling their products as meat alternatives. These products meet a number of important consumer wishes. They are, for example, healthy, easy-to-prepare (convenience) and show similarity with familiar foods (taste and texture) (Kuijer and Wielenga, 1999). Nowadays Quorn is perceived as a substitute for meat more than other meat alternatives like soy.

Kuijer and Wielenga (1999) expect that in the future, due to technological development, novel proteins will compete with meat on consumer aspects as price, taste and health.

3. Insects as food

Another interesting group of organism that can be used as a protein source are insects. In the Western world, insects as a food source are likely to

be met with scepticism, if not disgust. At the same time, insects are already used and accepted as food in many other countries. They are nutritious being a good source of protein, vitamins and energy (Ramos-Elorduy et al., 1997 and Bukkens, 1997).

At least 1386 edible insect species are consumed (Ramos-Elorduy et al., 1997). Insects are consumed at all stages of development: eggs, larvae, pupae and adults. They are included as a planned part of the diet throughout the year or when seasonally available as a side dish, snack or ingredient of dishes. Most eaten insects include grasshoppers, caterpillars, termites and aquatic insects (DeFoliart, 1992 and Bukkens, 1997). Eating insects is regarded as safe, since mankind has a long history of eating them. However, in the Western world they are not familiar as a food group. Nevertheless, the interest for insects as food in the Western world starts to increase considering more restaurants are serving insects as a delicacy and the amount of insect cookbooks is increasing. Actually, many of us already eat insects without realising it. For example, red scale insects are used as a colouring agent in Smarties, yoghurt, Campari and so on. Insects grown commercially for consumption include mealworms, crickets and wax moth larvae (DeFoliart, 1999). For example, cooked grasshoppers in can are $4 per 150 g.

4. Nutritional value of insects

Ramos-Elorduy and co-workers (1997) analysed several insects that are currently eaten. The conclusion was that the species provide high quality proteins and supplement the diet significantly with minerals and vitamins that are often in short supply in developing countries.

In general the protein content of insects ranges from 40 to 75 g/100 g dry weight, which is comparable to the protein content of meat (Bukkens, 1997 and Ramos-Elorduy et al., 1997). Most insect species convert plant protein to insect protein very efficiently. DeFoliart (1992) estimated that the food conversion efficiency of crickets is more than five times that of beef.

Table 1 provides information on the nutritional value of edible insects. The variation within an insect group is due to external factors, like climate, diet and habitat of the insects.

Table 1.
Nutritional value of insects (g/100 g dry weight)

	Protein	Fat	Mineral	Carbohydrates		Kcal
				Structural	C	
Orthoptera						
Grasshoppers, locusts	61–77	4–17	2–17	9–12	4	362–427
Coleoptera						
Beetles	21–54	18–52	1–7	6–23	1	410–574
Lepidoptera						
Butterflies, moths	15–60	7–77	3–8	2–29	1	293–762
Hymenoptera						
Bees, ants	1–81	4–62	0–6	1–6	8	416–655

	Protein	Fat	Mineral	Carbohydrates		Kcal
				Structural	C	
Meat[a]	45–55	40–57	1.4–2.3	0–1.5	0	433–652

Derived from **Bukkens (1997)**, **Ramos-Elorduy et al. (1997)**.

[a]The values for meat are derived from Nevo-tabel, original data in g/100 g product.

The quality of the protein, and thus the nutritional value, is determined by the amino acid composition and the digestibility of the proteins (Ladron de Guevara et al., 1995).

In Table 2 the amino acids per insect are shown, as well as the nutritional requirements for adults and preschoolers. Since children are in the growth phase their need for amino acids is higher in quantity. Besides, amino acids like histidine and arginine are more essential to them. Most insects contain sufficient amino acids to fulfil the nutritional requirements. The essential amino acid content score of insects ranges from 46 to 96% (Ramos-Elorduy et al., 1997). The first limiting essential amino acid in the majority of insects is either tryptophan or lysine (Bukkens, 1997 and Ramos-Elorduy et al., 1997).

Table 2.

Amino acid content of insects (g/100 g dry weight) and daily requirements for humans (mg/kg body weight per day)

Table2.

Amino acid content of insects (g/100 g dry weight) and daily requirements for humans (mg/kg body weight per day)

	Essential amino acids											Non-essential amino acids						
Order	Isolecine	Leucine	Lysine	Methionine	Cysteine[a]	Phenylalanine	Tyrosine[a]	Threonine	Tryptophan	Valine	Histidine	Aspartic acid	Serine	Glutamic acid	Proline	Glycine	Alanine	Arginine
Orthoptera																		
Grasshoppers and locusts	4.2–5.3	8.7–8.9	5.5–5.7	1.8–2.5	1.3–1.8	10.3–11.7	6.3–7.3	3.1–4.4	0.6–0.7	5.1–5.7	1.9–2.4	8.7–9.3	4.8–5.1	5.3–10.7	6.2–7.2	5.3–6.8	6.4–7.6	6.0–6.6
Coleoptera																		
Beetles	4.8–5.8	7.8–10.0	5.5–5.7	2.0–2.0	2.0–2.2	4.6–4.7	4.2–6.4	4.0–4.0	0.7–0.8	6.2–7.0	1.5–2.2	9.1–9.1	3.7–6.6	10.3–15.7	5.4–6.2	6.1–9.2	6.5–8.0	4.4–5.9
Lepidoptera																		
Butterflies and moths	4.1–5.1	6.9–8.0	4.9–6.3	2.1–2.6	1.3–5.4	6.4–9.5	4.4–9.5	3.8–4.7	0.4–0.6	4.8–6.1	1.6–2.9	8.7–10.7	3.8–6.2	10.5–13.9	5.6–7.3	5.1–6.2	5.6–6.8	5.7–6.8
Hymenoptera																		
Bees, wasps and ants	4.1–6.4	6.3–11.5	3.6–7.4	1.3–3.4	0.9–2.9	3.3–8.8	4.1–7.5	4.0–4.9	0.3–0.7	5.3–6.7	2.2–3.6	7.4–9.8	3.8–5.1	10.4–17.7	6.3–7.9	5.8–7.5	4.9–6.6	3.4–6.4
WHO/FAO																		
Preschooler	28	66	58	25[b]		63[c]		34	11	35	19							
Adult	13	19	16	17[b]		19[c]		9	5	13	16							

[a]Semi essential amino acids; cysteine can be derived from methionine and tyrosine from phenylalanine. [b]Methionine + cysteine. [c]Phenylalanine + tyrosine.

(adopted from **Ramos-Elorduy et al., 1997** and **Ladron de Guevara et al., 1995**).

The protein digestibility of insect protein is high, 77 to 98% (Ramos-Elorduy et al., 1997). The values for insects with an exoskeleton are on the lower level, due to chitin. If the outer skeleton is removed the digestibility increases and is comparable to meat (DeFoliart, 1992).

Insects vary widely in fat content and thus energy. The fat content of insects ranges between 7 and 77 g/100 g dry weight and the caloric value of insects varies between 293 and 762 kcal/100 g (Ramos-Elorduy et al., 1997). These values depend on the diet of insects and insect species. For instance worms, caterpillars and termites are known to contain more fat (Bukkens, 1997). The fatty acids in insects are similar to those in poultry and fish. According to DeFoliart (1992) some insects contain more essential fatty acids, like linoleic and/or linolenic acids, compared with meat.

Overall the mineral content of insects ranges from 3 to 8 g/100 g dried sample (Ramos-Elorduy et al., 1997). They contain a high amount of zinc and iron (DeFoliart, 1992), which is more than beef (Bukkens, 1997). The calcium concentration is around 920 mg/100 g dry weight (Bukkens, 1997).

In conclusion, insects contain high quality proteins and supplement the diet significantly with minerals. Besides, they have a long history of safe consumption.

However, there are some disadvantages with regard to the production of insects as food. As mentioned before, the composition, and thus the nutritional value, of insects varies. Furthermore, the scaling up of insect breeding can be difficult. Not all species are suitable to culture at a large scale (DeFoliart, 1999). Some of the problems of breeding cattle may also occur at the insect farms, like increasing vulnerability to diseases and animal welfare. Another problem is the fact that insects can produce defensive secretions and/or can be a source of inhalant allergens (e.g. cast skins, excreta), which might be irritating for people working with them (DeFoliart, 1992).

5. Insect cells from bioreactors

Insect cells might be a good alternative for insects. They can be cultured in a bioreactor in suspension. Growth in bioreactors has several advantages. The conditions in the bioreactor are controllable resulting in a product of reproducible quality. Furthermore, the bioreactor is a closed system resulting in a low risk of contamination compared with breeding whole insects. Like whole insects, insect cells do not support growth of viruses or expression of oncogenes affecting humans (Agathos et al., 1990).

Finally, there are several methods that possibly can be used to rationally engineer the composition of the insect-cell biomass. First of all, the natural variation in composition between cell-lines derived from different insects may be used. In addition, insect cells derived from a different type of tissue may differ in composition, e.g. ovarian cells versus haemolymph cells. Secondly, it may be possible to influence the composition of the biomass by changing the process conditions, like medium composition, pH and temperature. Last, using modern biotechnology, the nutritional value of the insect-cells can be altered by infecting the cells with a recombinant baculovirus. At present baculoviruses are used in order to produce eukaryotic proteins, e.g. pharmaceuticals. Baculoviruses are not dangerous for humans and can infect insect cells. At the end of infection up to 50% of the protein in the cell can be derived from baculoviruses. The major part of the baculovirus

proteins is formed by only two proteins, being the P10 and the polyhedron protein. Thus, to enrich insect cells with essential amino acids a recombinant baculovirus could be used, where the polyhedron and P10 genes are replaced by genes that encode for a protein enriched for certain limiting essential amino acids. For example, a baculovirus with LacZ (a gene coding for beta-galactosidase) could in theory enrich insect cells with tryptophan.

6. Consumer point of view

In order to put a successful novel protein on the market it is important to meet consumer wishes like convenience, healthy diet and similarity with familiar food. Insect (biomass) is not a common source of food in Western countries. The question is how consumers perceive insects. Do we perceive insects as edible and does this perception change if the product contains insect proteins instead of whole insects. Does it matter whether these proteins come from whole insects or from insect cells out of a bioreactor and what if they are made using modern biotechnology? An additional question is whether consumer perception can be influenced. For example, the mycoprotein of Quorn is referred to as “a nutritious member of the fungi family, as are mushrooms and truffles”. Bringing these types of food to the market asks for an integrated approach of technological research and consumer behaviour studies to prevent preliminary rejection.

7. Concluding remarks

Insects hold potential as a safe, nutritious, flexible and reliable protein source for the future. They are nutritious in terms of protein content, essential amino acids, essential fatty acids and other micro-nutrients. In principle the same holds for insect cells, which in addition would allow for the rational engineering of the composition and thus nutritional value. Whether this actually can be done technologically and whether the nutritional value of insect cells is the same as the one of insects still needs to be studied. Besides technological aspects, a major point that should be studied is whether consumers will accept foods derived from insects and insect cells and whether this can be influenced using a proper marketing strategy.

Acknowledgement

The authors gratefully acknowledge Monique van Oers of the Laboratory of Virology for useful discussions on the application of recombinant baculoviruses.

Piloto 1B –

Responda as questões abaixo em português.

1) Qual é o objetivo da pesquisa?(1,3)

2) O que motivou a pesquisa? Explique.(1,3)

3) De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, preencha a tabela abaixo. (1,2)

Proteína	Sucesso ou Fracasso	Razão

4) Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento. (1,2)

5) Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína? (1,3)

6) Quais as vantagens e desvantagens da utilização da nova proteína como alimento? (1,2)

7) Os autores do artigo chegaram a que conclusão? (1,3)

8) Qual é o papel do consumidor nessa história? (1,2)

Teste Final –

Responda as questões abaixo em português.

1) Qual é o objetivo da pesquisa?

2) O que motivou a pesquisa? Explique.

3) De acordo com a perspectiva histórica de introdução de novas proteínas no mercado, escreva abaixo:
   c) a denominação das 3(três) proteínas que foram apresentadas ao mercado consumidor;
   d) a reação do mercado consumidor (se for apresentada) e a razão do sucesso ou fracasso de cada uma delas.

4) Qual é a diferença apresentada no item 3 do texto, quanto à aceitação da nova fonte de proteína como alimento.

5) Qual a conclusão apresentada sobre o valor nutricional da nova fonte de proteína?

6) Quais as vantagens e desvantagens da utilização da nova proteína como alimento?
   Vantagens:

   Desvantagens:

7) Diante da dificuldade que os consumidores demonstram em incluir os insetos em seus cardápios, os pesquisadores pensam em uma outra possibilidade. Diga qual a possibilidade e explique por que eles a consideram viável.

8) Que argumentos podem justificar mais estudos em relação ao ponto de vista do consumidor?

PILOTO 1C

(PILOTO 1D **- mesmo texto / questões diferentes)**

Sufficiency Test – Pilot 1C

Student: _____

Institution: _____ **Course:** _____

Teacher: Graziella Lapkoski - UFPR **Grade:**_____ **Nov. 07**

***Permitido o uso do dicionário!**

****Teste deve ser respondido em português!** **Usar caneta azul ou preta!**

Tempo de duração previsto: entre 1:30h e 2:00h.

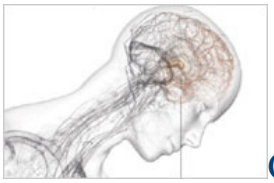
In Diabetes, a Complex of Causes

By AMANDA SCHAFFER

Published: October 16, 2007

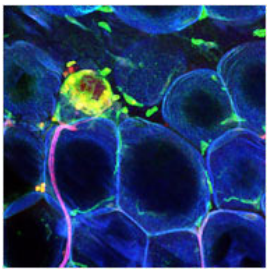
An explosion of new research is vastly changing scientists' understanding of diabetes and giving new clues about how to attack it.

Multimedia



Graphic

The Body's Role in Diabetes



Life Sciences Institute/ University of Michigan

Study Fat tissue from a mouse that was fed a high-fat diet.

The fifth leading killer of Americans, with 73,000 deaths a year, diabetes is a disease in which the body's failure to regulate glucose, or blood sugar, can lead to serious and even fatal complications. Until very recently, the regulation of glucose — how much sugar is present in a person's blood, how much is taken up by cells for fuel, and how much is released from energy stores — was regarded as a conversation between a few key players: the pancreas, the liver, muscle and fat.

Now, however, the party is proving to be much louder and more complex than anyone had shown before.

New research suggests that a hormone from the skeleton, of all places, may influence how the body handles sugar. Mounting evidence also demonstrates that signals from the immune system, the brain and the gut play critical roles in controlling glucose and lipid metabolism. (The findings are

mainly relevant to Type 2 diabetes, the more common kind, which comes on in adulthood.)

Focusing on the cross-talk between more different organs, cells and molecules represents a "very important change in our paradigm" for understanding how the body handles glucose, said Dr. C. Ronald Kahn, a diabetes researcher and professor at Harvard Medical School.

The defining feature of diabetes is elevated blood sugar. But the reasons for abnormal sugar seem to "differ tremendously from person to person," said Dr. Robert A. Rizza, a professor at the Mayo Clinic College of Medicine. Understanding exactly what signals are involved, he said, raises the hope of "providing the right care for each person each day, rather than giving everyone the same drug."

Last summer, researchers at Columbia University Medical Center published startling results showing that a hormone released from bone may help regulate blood glucose.

When the lead researcher, Dr. Gerard Karsenty, first described the findings at a conference, the assembled scientists "were overwhelmed by the potential implications," said Dr. Saul Malozowski, senior adviser for endocrine physiology research at the National Institute of Diabetes and Digestive and Kidney Diseases, who was not involved in the research. "It was coming from left field. People thought, 'Oof, this is really new.'

"For the first time," he went on, "we see that the skeleton is actually an endocrine organ," producing hormones that act outside of bone.

In previous work, Dr. Karsenty had shown that leptin, a hormone produced by fat, is an important regulator of bone metabolism. In this work, he tested the idea that the conversation was a two-way street. "We hypothesized that if fat regulates bone, bone in essence must regulate fat," he said.

Working with mice, he found that a previously known substance called osteocalcin, which is produced by bone, acted by signaling fat cells as well as the pancreas. The net effect is to improve how mice secrete and handle insulin, the hormone that helps the body move glucose from the bloodstream into cells of the muscle and liver, where it can be used for energy or stored for future use. Insulin is also important in regulating lipids.

In Type 2 diabetes, patients' bodies no longer heed the hormone's directives. Their cells are insulin-resistant, and blood glucose levels surge. Eventually, production of insulin in the pancreas declines as well.

Dr. Karsenty found that in mice prone to Type 2 diabetes, an increase in osteocalcin addressed the twin problems of insulin resistance and low insulin production. That is, it made the mice more sensitive to insulin and it increased their insulin production, thus bringing their blood sugar down. As a bonus, it also made obese mice less fat.

If osteocalcin works similarly in humans, it could turn out to be a "unique new treatment" for Type 2 diabetes, Dr. Malozowski said. (Most current diabetes drugs either raise insulin production or improve insulin sensitivity, but not both. Drugs that increase production tend to make insulin resistance worse.)

A deficiency in osteocalcin could also turn out to be a cause of Type 2 diabetes, Dr. Karsenty said. Another recent suspect in glucose regulation is the immune system. In 2003, researchers from two laboratories found that fat tissue from obese mice contained an abnormally large number of macrophages,

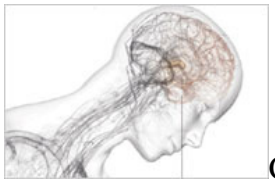
immune cells that contribute to inflammation. The finding piqued the curiosity of researchers. “I remember reading the paper and thinking: ‘Wow, look at all those macrophages. What are they doing?’” said Dr. Jerrold M. Olefsky of the University of California, San Diego, School of Medicine.

Scientists have long suspected that inflammation was somehow related to insulin resistance, which precedes nearly all cases of Type 2 diabetes. In the early 1900s, diabetics were sometimes given high doses of aspirin, which is an anti-inflammatory, Dr. Olefsky said.

Only in the past few years has research into the relationship of obesity, inflammation and insulin resistance become “really hot,” said Dr. Alan R. Saltiel, director of the Life Sciences Institute at the University of Michigan.

Many researchers agree that obesity is accompanied by a state of chronic, low-grade inflammation in which some immune cells are activated, and that that may be a primary cause of insulin resistance. They also agree that the main type of cell responsible for the inflammation is the macrophage, Dr. Saltiel said.

Multimedia



Graphic

The Body’s Role in Diabetes

But major questions remain, he said: “Why are these macrophages attracted to fat, liver and muscle in the first place? What are they doing? What are they secreting? What other immune cells are in there?”

New research also suggests that “not all macrophages are created equal,” added Dr. Saltiel. There appear to be “good ones and bad ones” competing in fat tissue, with potentially large consequences for inflammation and diabetes.

Meanwhile, the promise of anti-inflammatory compounds as treatment continues to attract attention. “Certain cellular anti-inflammatory proteins may now be important new targets for drug discovery for diabetes treatment,” Dr. Olefsky said. But damping down the immune system is also potentially risky, he noted, adding: “If you’re inhibiting the macrophage inflammatory pathway, that’s good for insulin resistance and diabetes. But it might not be so good for your susceptibility to infections.” A major goal is to develop a drug that quashes only the specific component of macrophage inflammation that leads to insulin resistance, without causing other side effects.

One class of current medications, called thiazolidinediones, may work in part by reducing inflammation, which may in turn improve insulin sensitivity. But an example from this class, the drug Avandia, was also found to increase the risk of heart attacks.

Another participant in the glucose conversation is the brain. Its role has long been suspected. More than a century ago, the French physiologist Claude Bernard suggested that the brain was important in blood sugar regulation. He punctured the brains of experimental animals in specific areas and managed to derange their blood sugar metabolism, making them diabetic.

But for years, virtually no one followed up on this finding, said Dr. Kahn, of Harvard.

People thought about glucose as a critical fuel for the brain, Dr. Kahn said, but did not explore the brain's role in glucose regulation.

Only recently, with more advanced laboratory techniques, has this role been definitively established and expanded upon.

Today's genetic techniques, said Dr. Rizza, at the Mayo Clinic, are what have "really driven the process."

For instance, once scientists developed the ability to manipulate mice so that they lacked particular receptors in specific tissues, they could show that mice without insulin receptors in the brain could not regulate glucose properly and went on to develop diabetes, said Dr. Kahn, whose laboratory published this groundbreaking work in 2000.

Other researchers have shown that free fatty acids, as well as the hormone leptin, produced by fat tissue, signal directly to a part of the brain called the hypothalamus, which also regulates appetite, temperature and sex drive.

And several recent papers suggest that direct signaling by glucose itself to neurons in the hypothalamus is also crucial to normal blood sugar regulation in mice.

"If the brain is getting the message that you have adequate amounts of these hormones and nutrients, it will constrain glucose production by the liver and keep blood glucose relatively low," said Dr. Michael W. Schwartz, a professor at the University of Washington. But if the brain senses inadequate amounts, he continued, it will "activate responses that cause the liver to make more glucose, and new evidence suggests that this contributes to diabetes and impaired glucose metabolism."

The brain, therefore, appears to be listening to — and weighing and making sense of — a chorus of signals from insulin, leptin, free fatty acids and glucose itself. In response, it appears to send signals to liver and muscle cells by way of several nerves, though additional mechanisms are probably involved. The gut also seems to chime in, said Dr. Rizza, adding that for him, this aspect of sugar regulation came as "the biggest gee whiz of all."

"Food comes in through the gut, so of course you should look there" for molecules involved in glucose regulation, he said. "But few people realized this until very recently."

Hormones from the small intestine called incretins turn out to talk directly with the brain and pancreas in ways that help reduce blood sugar and cause animals and people to eat less and lose weight, Dr. Rizza said.

Numerous molecules that mimic incretins or prevent them from being degraded are in clinical trials. Two such drugs have been approved by the Food and Drug Administration: Byetta, an incretin mimic, from Amylin Pharmaceuticals and Eli Lilly; and Januvia, from Merck, which inhibits the destruction of the incretin GLP1. (Dr. Rizza is an adviser to Merck but says all consulting fees go to the Mayo Clinic for education and research.)

Still, it can be hard to predict how different drugs will interact in the body. And many promising candidates will turn out to have side effects — chattering helpfully with one organ, but problematically with another.

"The picture is becoming more and more complicated," Dr. Saltiel said. "And let's face it, it was pretty complicated before."

Responda as questões abaixo em português. (2,0 x 5=10,0)

1) Quais os objetivos da pesquisa apresentados no texto? (Explícitos e/ou implícitos)

2) Quais as hipóteses levantadas nas diferentes pesquisas mencionadas no texto?

3) O que é dito em relação ao tratamento da doença?

4) Quais foram os resultados obtidos nas pesquisas?

5) A que conclusão a autora do texto chega?

PILOTO 1D

(mesmo texto do PILOTO 1C **- questões diferentes)**

Sufficiency Test – Pilot 1D

Student: _____

Institution: _____ **Course:** _____

Teacher: Graziella Lapkoski - UFPR **Grade:**_____ **Nov. 07**

***Permitido o uso do dicionário!**

****Teste deve ser respondido em português!** ***Usar caneta azul ou preta!***

Tempo de duração previsto: entre 1:30h e 2:00h.

Responda as questões abaixo em português. (1,0 x 10=10,0)

1) A autora do texto diz que: “Now, however, the party is proving to be much louder and more complex than anyone had shown before.” Explique o que ela quiz dizer com isso.

2) Quais as novas descobertas sobre diabetes?

3) que essas descobertas significam em termos de tratamentos?

4) Que idéia de um antigo pesquisador serviu de motivação para a pesquisa?

5) Qual a diferença entre as drogas já existentes e uma, possivelmente desenvolvida à partir dos resultados da pesquisa realizada?

6) Que papel o esqueleto desempenha na questão do diabetes?

7) Qual é o papel desempenhado pelo cérebro?

8) Como os cientistas chegaram à conclusão relativa ao papel do cérebro?

9) Que outro órgão, mencionado no final do texto, também está envolvido na regulagem do açúcar? Como?

10) Essas descobertas e o desenvolvimento das novas drogas são conclusivas? Por que?

3– TABELAS – *THE SUMMARY STATISTICS* – PILOTO 1B E TESTE FINAL

The Summary Statistics – Pilot 1B

```
RUMM2020        Project: PILOTO1BB        Analysis: PILOTO1B
Title:  PILOTO1B                             Date: 7 out 2008 01:33:41
Display: SUMMARY TEST-OF-FIT STATISTICS
                      ITEM-PERSON INTERACTION
=============================================================================
                          ITEMS                          PERSONS
                 Location  Fit Residual      Location  Fit Residual
-----------------------------------------------------------------------------
Mean                0,000      0,578            0,254      0,274
SD                  0,789      0,722            0,320      0,799
Skewness                       0,024                       0,425
Kurtosis                      -1,183                      -0,938
Correlation                   -0,015                       0,144

Complete data DF =             0,782
-----------------------------------------------------------------------------
=============================================================================
        ITEM-TRAIT INTERACTION            RELIABILITY INDICES
-----------------------------------------------------------------------------
Total Item Chi Squ           49,555     Separation Index0,79341
Total Deg of Freedom         32,000     Cronbach Alpha      N/A
Total Chi Squ Prob          0,024592
-----------------------------------------------------------------------------
=============================================================================
        LIKELIHOOD-RATIO TEST             POWER OF TEST-OF-FIT
-----------------------------------------------------------------------------
Chi Squ                                  Power is GOOD
Degrees of Freedom                       [Based on SepIndex of 0,79341]
Probability
-----------------------------------------------------------------------------
```

The Summary Statistics – Final Test

```
RUMM2020        Project: TFINAL        Analysis: TESTEF
Title:  TESTE FINAL                          Date: 7 out 2008 01:37:33
Display: SUMMARY TEST-OF-FIT STATISTICS

                        ITEM-PERSON INTERACTION
=============================================================================
                          ITEMS                          PERSONS
                 Location  Fit Residual      Location  Fit Residual
-----------------------------------------------------------------------------
Mean                0,000      0,163           -0,668     -0,061
SD                  0,664      0,992            0,897      0,981
Skewness                       0,778                      -0,223
Kurtosis                      -0,143                       0,287
Correlation                   -0,152                       0,093

Complete data DF =             0,871
-----------------------------------------------------------------------------
=============================================================================
        ITEM-TRAIT INTERACTION            RELIABILITY INDICES
-----------------------------------------------------------------------------
Total Item Chi Squ           89,027     Separation Index0,91017
Total Deg of Freedom         64,000     Cronbach Alpha      N/A
Total Chi Squ Prob          0,021027
-----------------------------------------------------------------------------
=============================================================================
        LIKELIHOOD-RATIO TEST             POWER OF TEST-OF-FIT
-----------------------------------------------------------------------------
Chi Squ                                  Power is EXCELLENT
Degrees of Freedom                       [Based on SepIndex of 0,91017]
Probability
-----------------------------------------------------------------------------
```

4– EXEMPLO DA TABELA DE DADOS PARA ALIMENTAÇÃO DO APLICATIVO

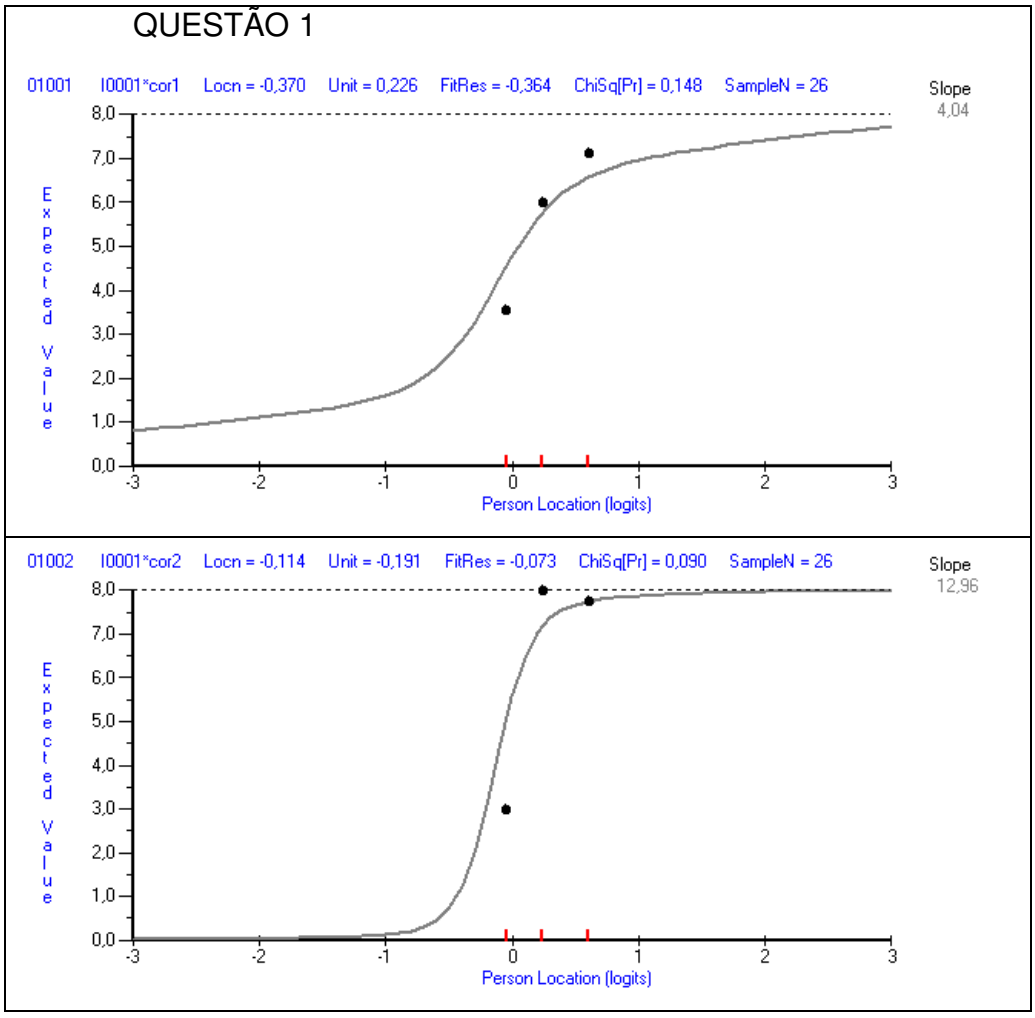
```
001121342230
001322310231
002201400110
002301401310
003222541101
003322442201
004210301313
004320200533
005211012000
005443003002
006100100000
006200110100
007114411311
007215320211
008124214210
008204323111
009110123200
009221233100
010152234330
010342223240
011151111230
011351202340
012205100000
012315100000
013130300200
013330300200
014223312100
014433414100
015210100000
015300100000
016233533452
016444434452
017130421223
017232321123
018224555034
018344545145
019221311000
019322311100
020102340000
020314410000
```

5– GRÁFICOS DAS CURVAS CARACTERÍSTICAS DOS ITENS PILOTO 1B E TESTE FINAL

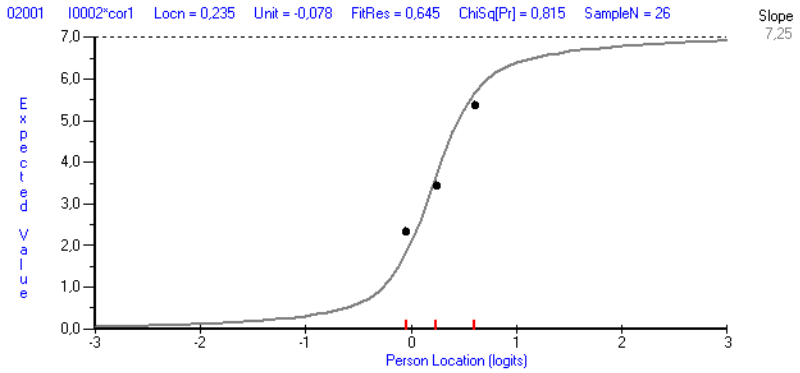
TESTE 1B

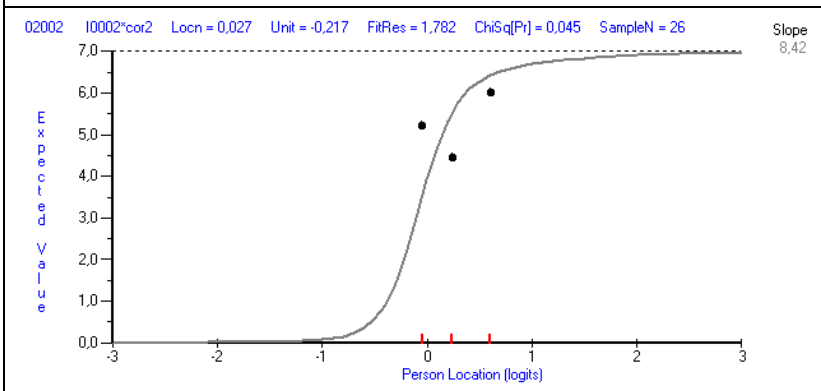
CURVAS CARACTERÍSTICAS DOS ITENS (ITEMS CHARACTERISTIC CURVES)

QUESTÃO 1

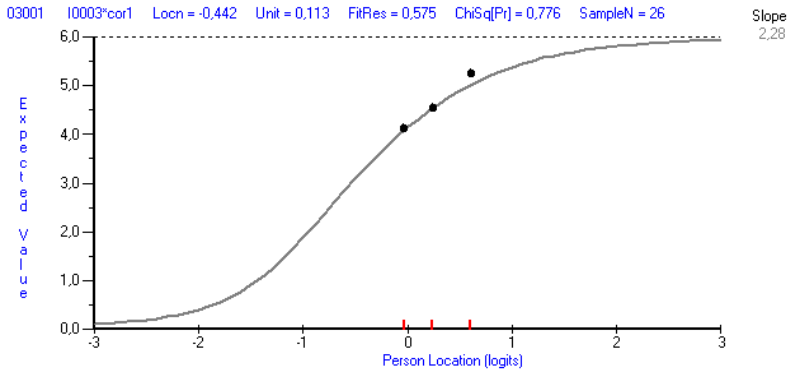


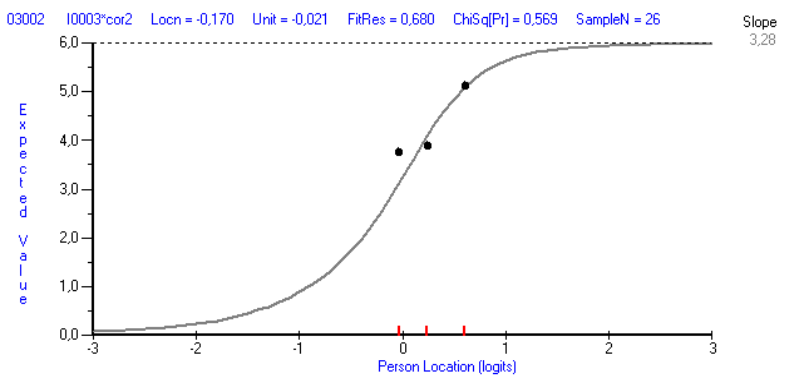
QUESTÃO 2



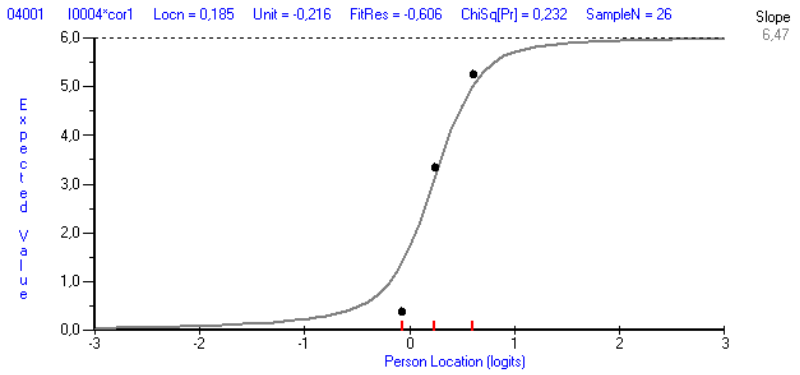


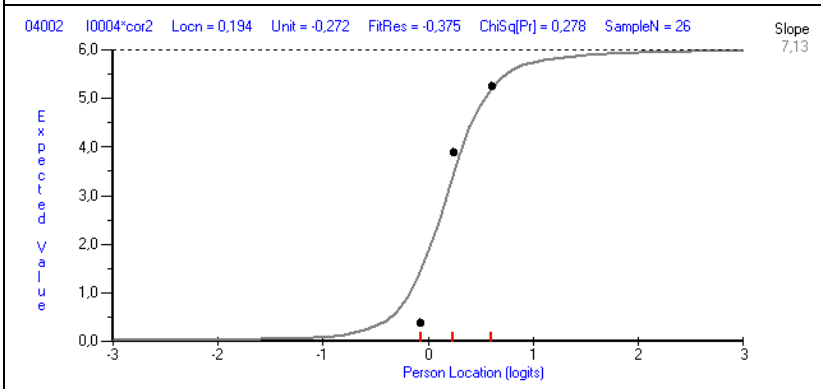
QUESTÃO 3



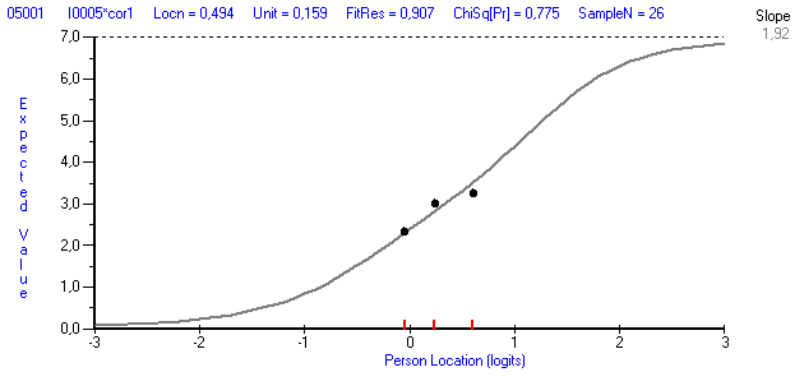


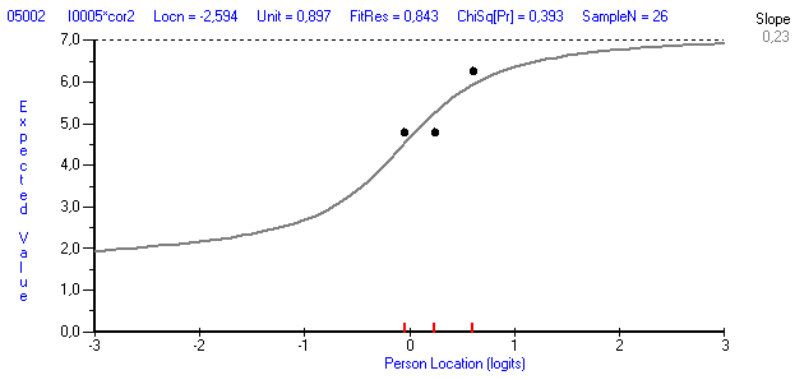
QUESTÃO 4



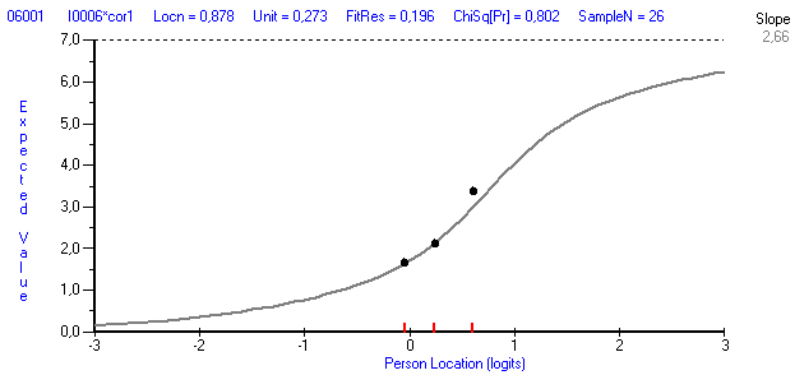


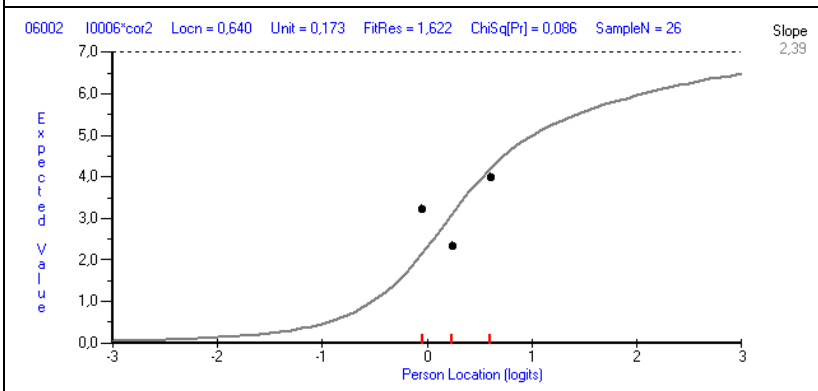
QUESTÃO 5



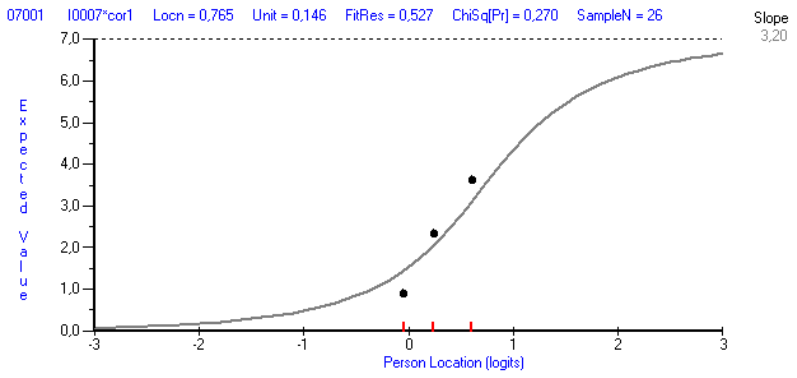


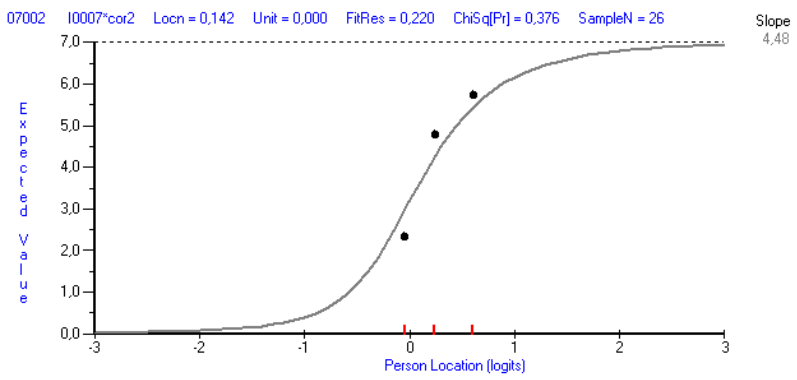
QUESTÃO 6



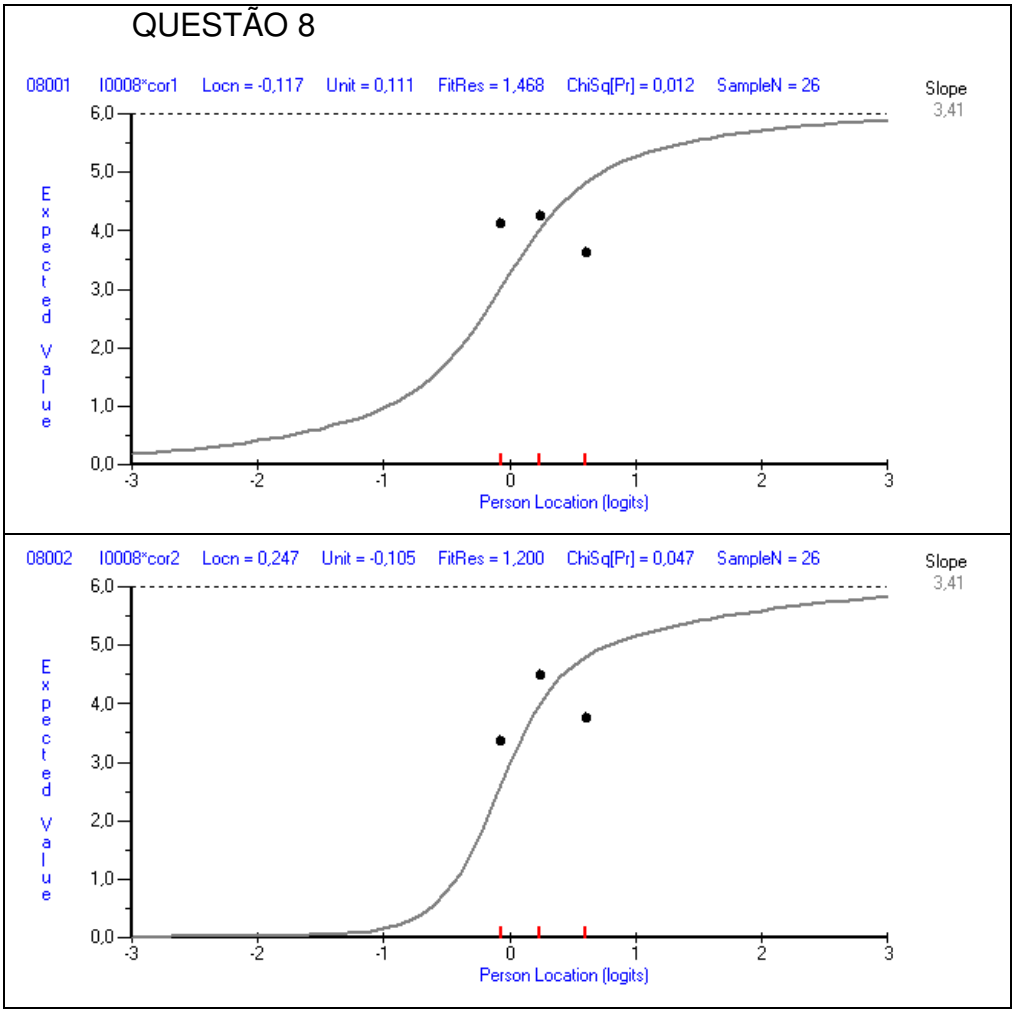


QUESTÃO 7





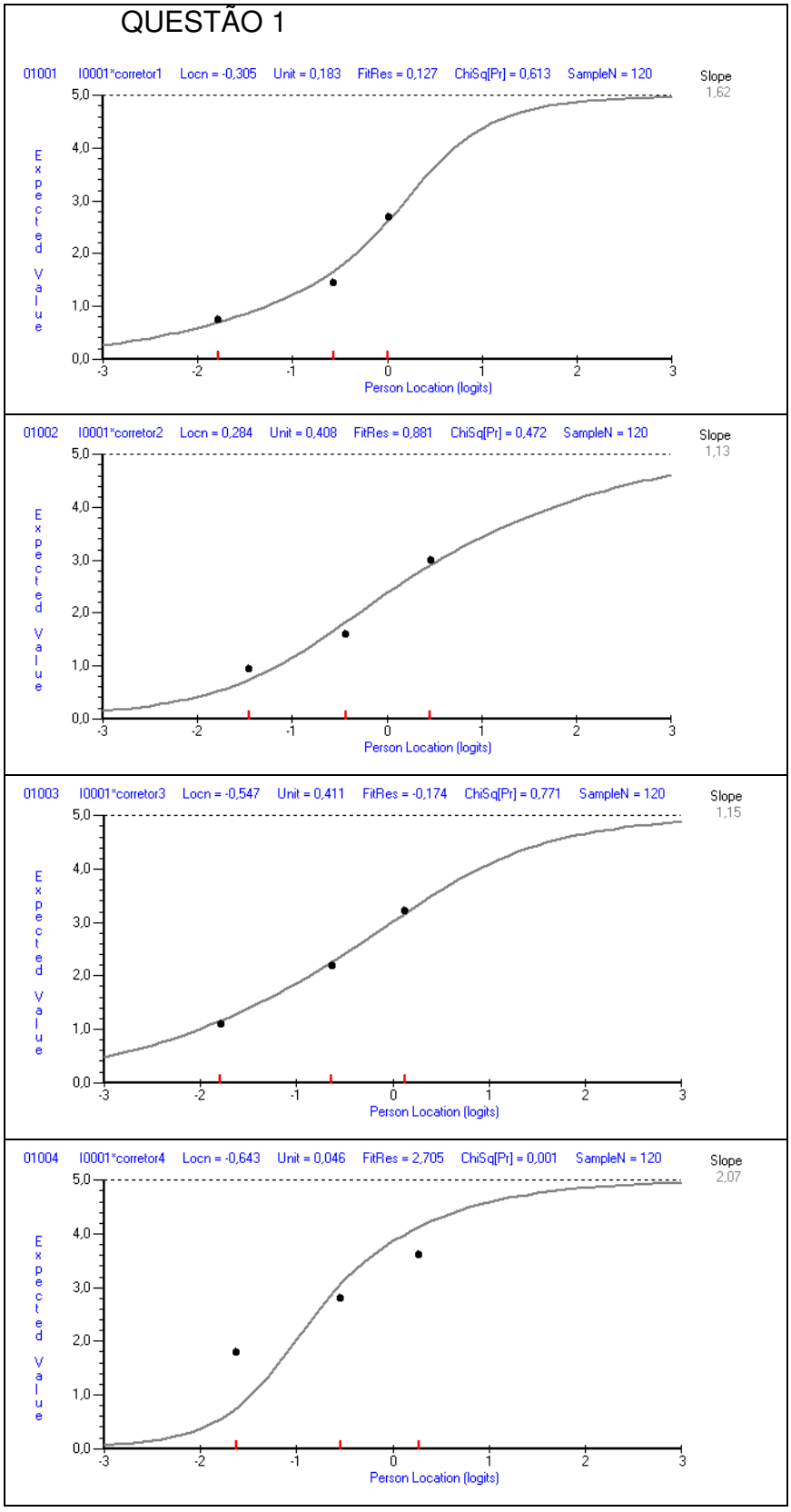
QUESTÃO 8


08001 I0008*cor1 Locn = -0,117 Unit = 0,111 FitRes = 1,468 ChiSq[Pr] = 0,012 SampleN = 26
Slope
3,41
Expected Value
6,0
5,0
4,0
3,0
2,0
1,0
0,0
-3
-2
-1
0
1
2
3
Person Location (logits)
08002 I0008*cor2 Locn = 0,247 Unit = -0,105 FitRes = 1,200 ChiSq[Pr] = 0,047 SampleN = 26
Slope
3,41
Expected Value
6,0
5,0
4,0
3,0
2,0
1,0
0,0
-3
-2
-1
0
1
2
3
Person Location (logits)

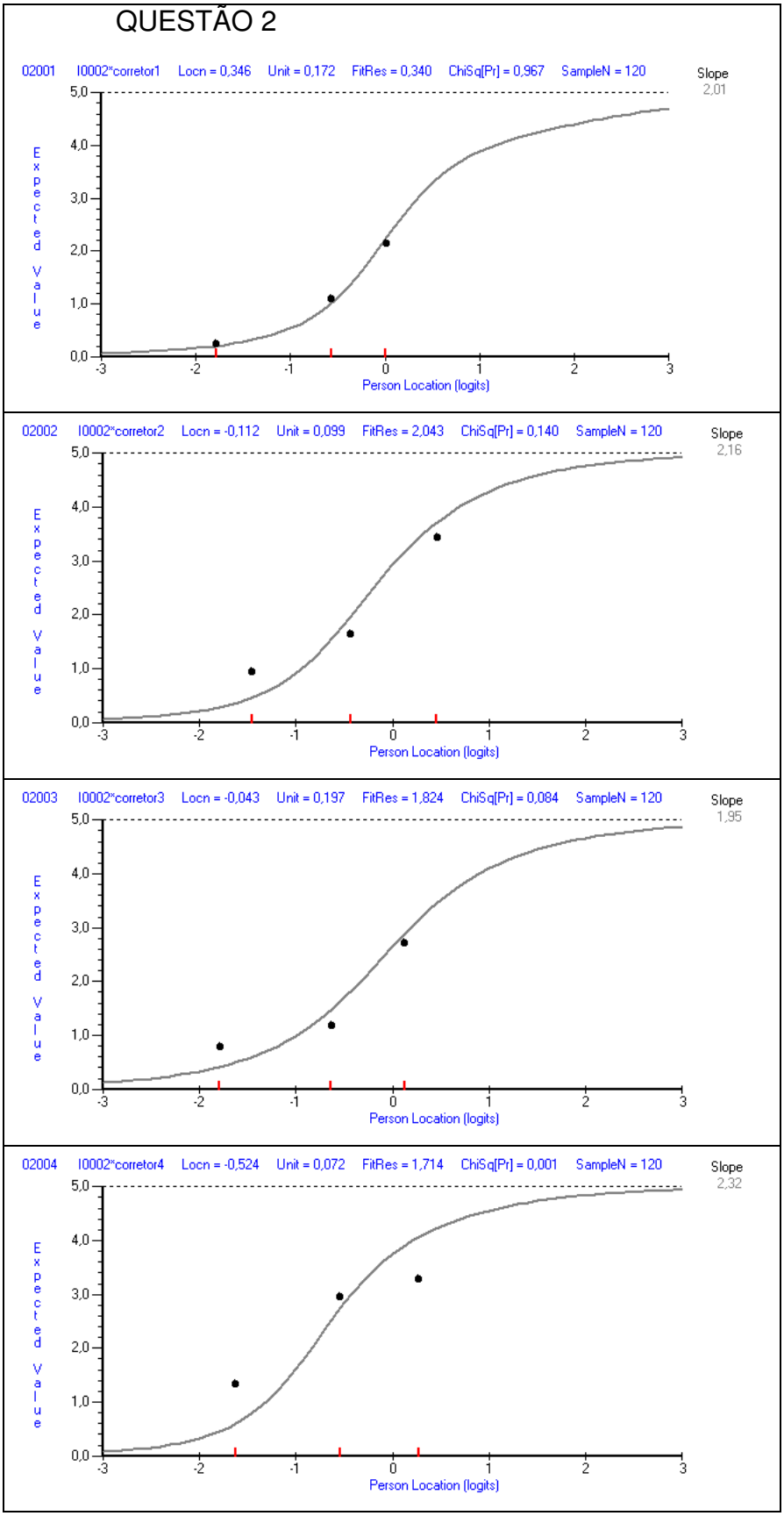
TESTE FINAL

CURVAS CARACTERÍSTICAS DOS ITENS (ITEMS CHARACTERISTIC CURVES)

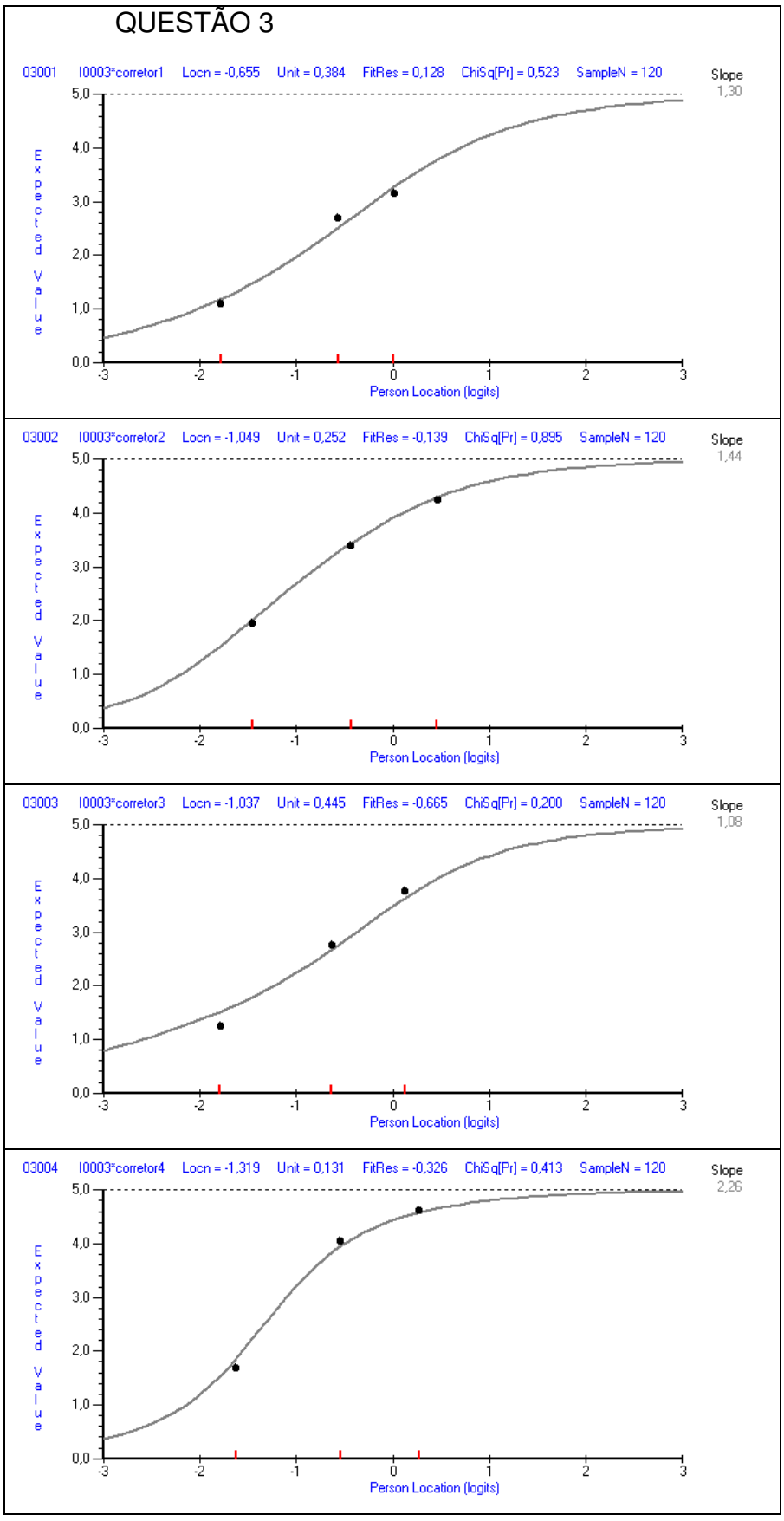
QUESTÃO 1



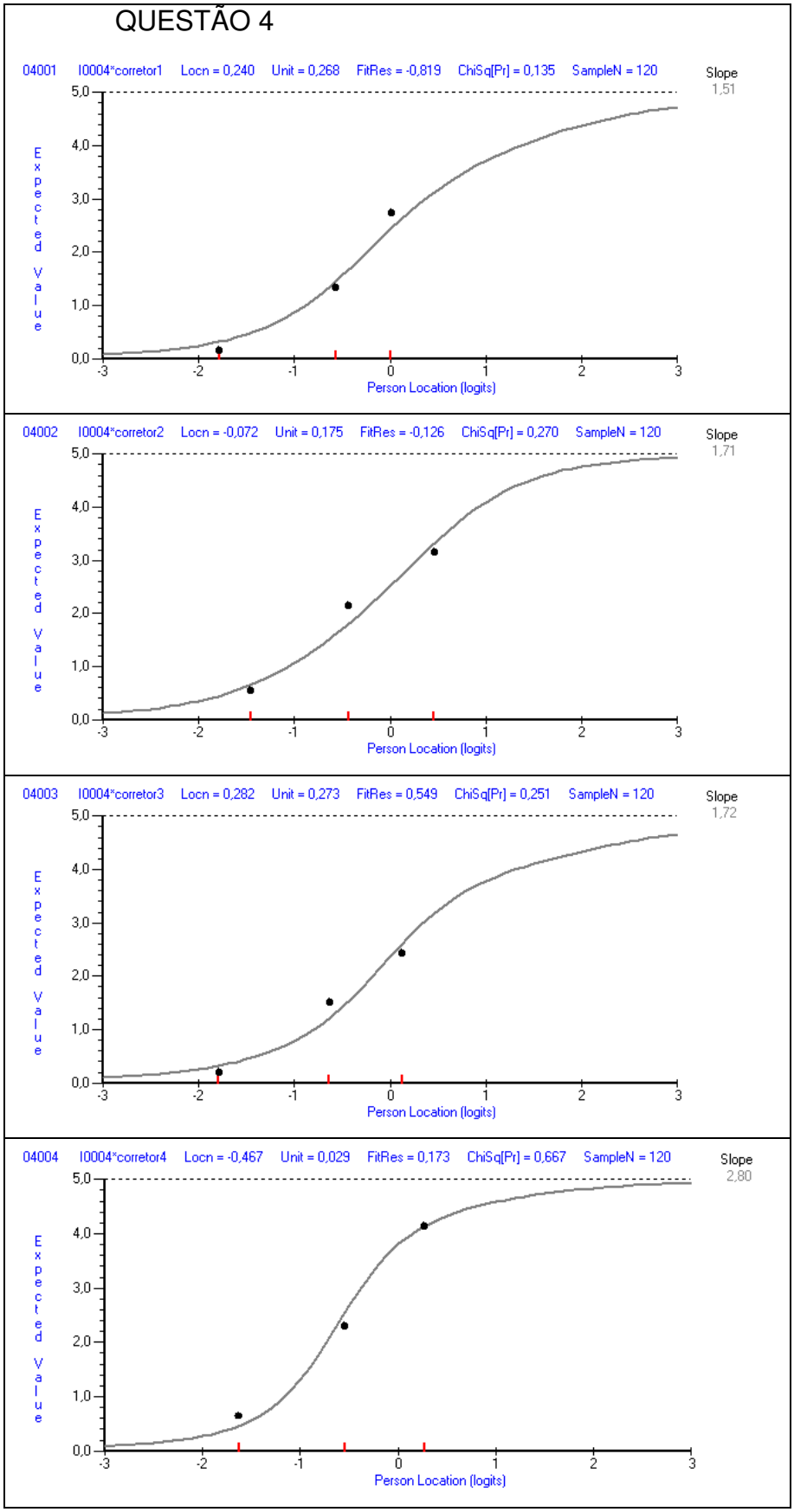
QUESTÃO 2


02001 I0002*corretor1 Locn = 0,346 Unit = 0,172 FitRes = 0,340 ChiSq[Pr] = 0,967 SampleN = 120
Slope 2,01
02002 I0002*corretor2 Locn = -0,112 Unit = 0,099 FitRes = 2,043 ChiSq[Pr] = 0,140 SampleN = 120
Slope 2,16
02003 I0002*corretor3 Locn = -0,043 Unit = 0,197 FitRes = 1,824 ChiSq[Pr] = 0,084 SampleN = 120
Slope 1,95
02004 I0002*corretor4 Locn = -0,524 Unit = 0,072 FitRes = 1,714 ChiSq[Pr] = 0,001 SampleN = 120
Slope 2,32
Expected Value
Person Location (logits)

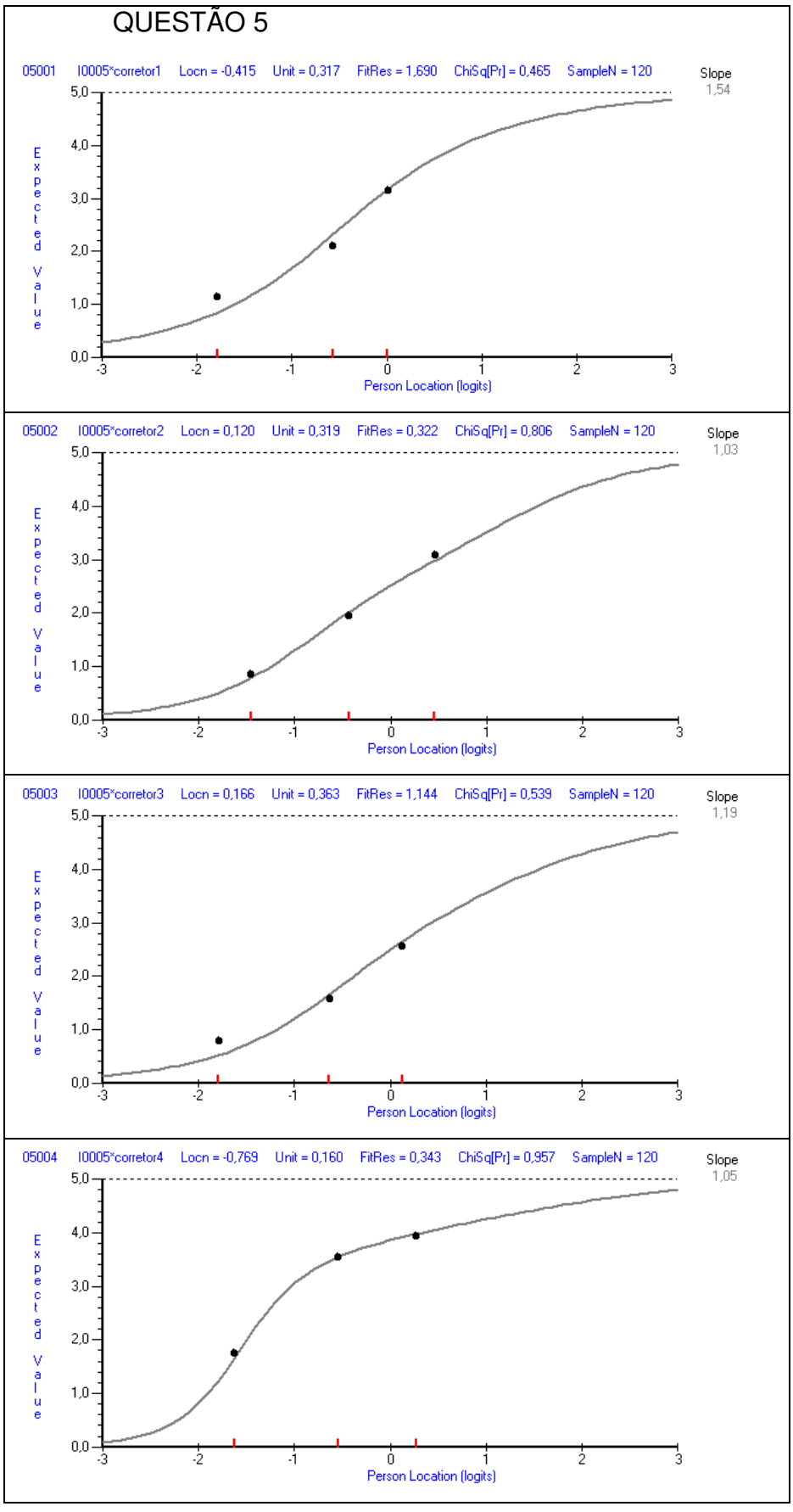
QUESTÃO 3


03001 I0003*corretor1 Locn = -0,655 Unit = 0,384 FitRes = 0,128 ChiSq[Pr] = 0,523 SampleN = 120
Slope 1,30
03002 I0003*corretor2 Locn = -1,049 Unit = 0,252 FitRes = -0,139 ChiSq[Pr] = 0,895 SampleN = 120
Slope 1,44
03003 I0003*corretor3 Locn = -1,037 Unit = 0,445 FitRes = -0,665 ChiSq[Pr] = 0,200 SampleN = 120
Slope 1,08
03004 I0003*corretor4 Locn = -1,319 Unit = 0,131 FitRes = -0,326 ChiSq[Pr] = 0,413 SampleN = 120
Slope 2,26
Expected Value
Person Location (logits)

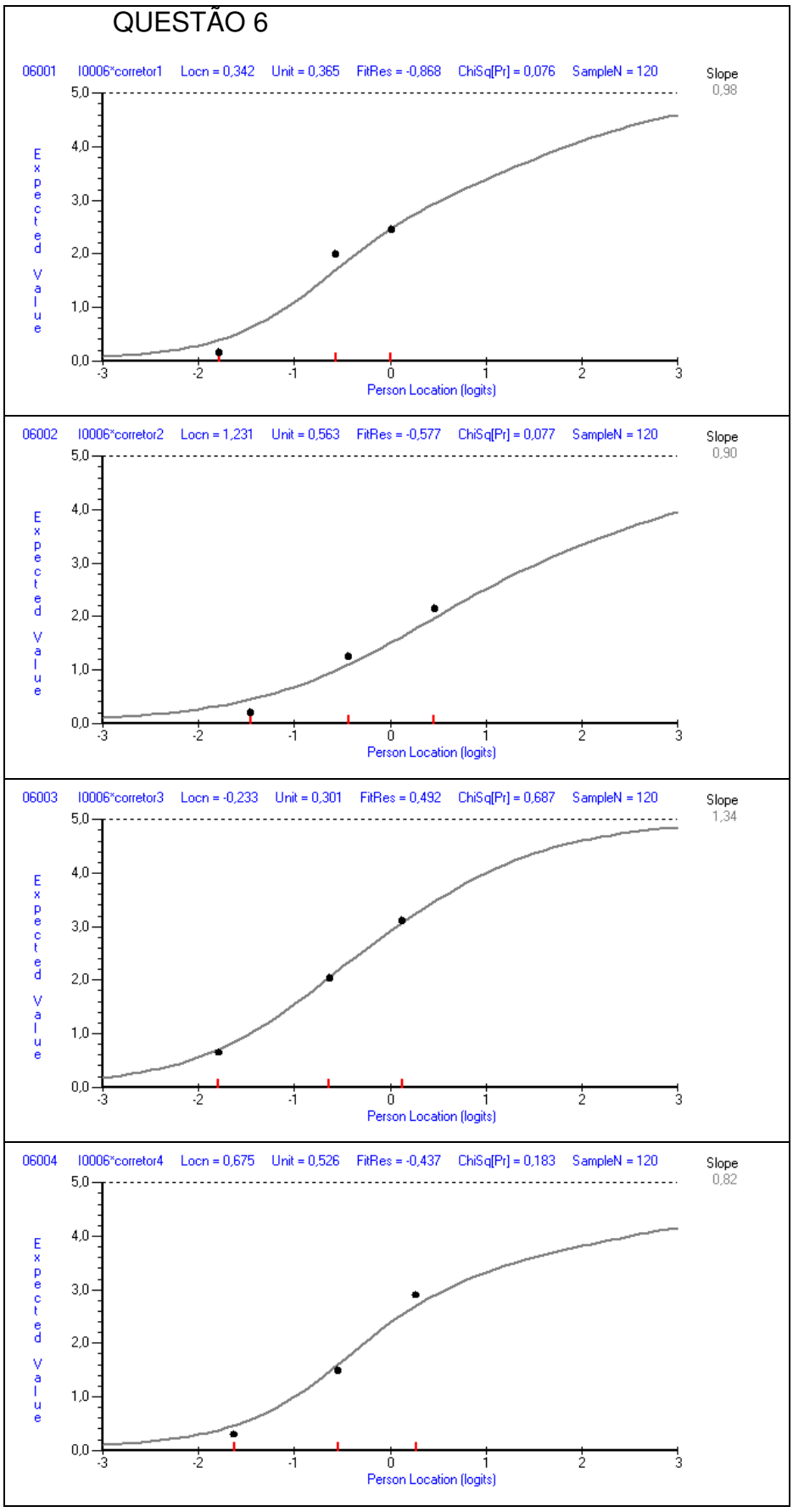
QUESTÃO 4


04001 I0004*corretor1 Locn = 0,240 Unit = 0,268 FitRes = -0,819 ChiSq[Pr] = 0,135 SampleN = 120
Slope 1,51
04002 I0004*corretor2 Locn = -0,072 Unit = 0,175 FitRes = -0,126 ChiSq[Pr] = 0,270 SampleN = 120
Slope 1,71
04003 I0004*corretor3 Locn = 0,282 Unit = 0,273 FitRes = 0,549 ChiSq[Pr] = 0,251 SampleN = 120
Slope 1,72
04004 I0004*corretor4 Locn = -0,467 Unit = 0,029 FitRes = 0,173 ChiSq[Pr] = 0,667 SampleN = 120
Slope 2,80
Expected Value
Person Location (logits)

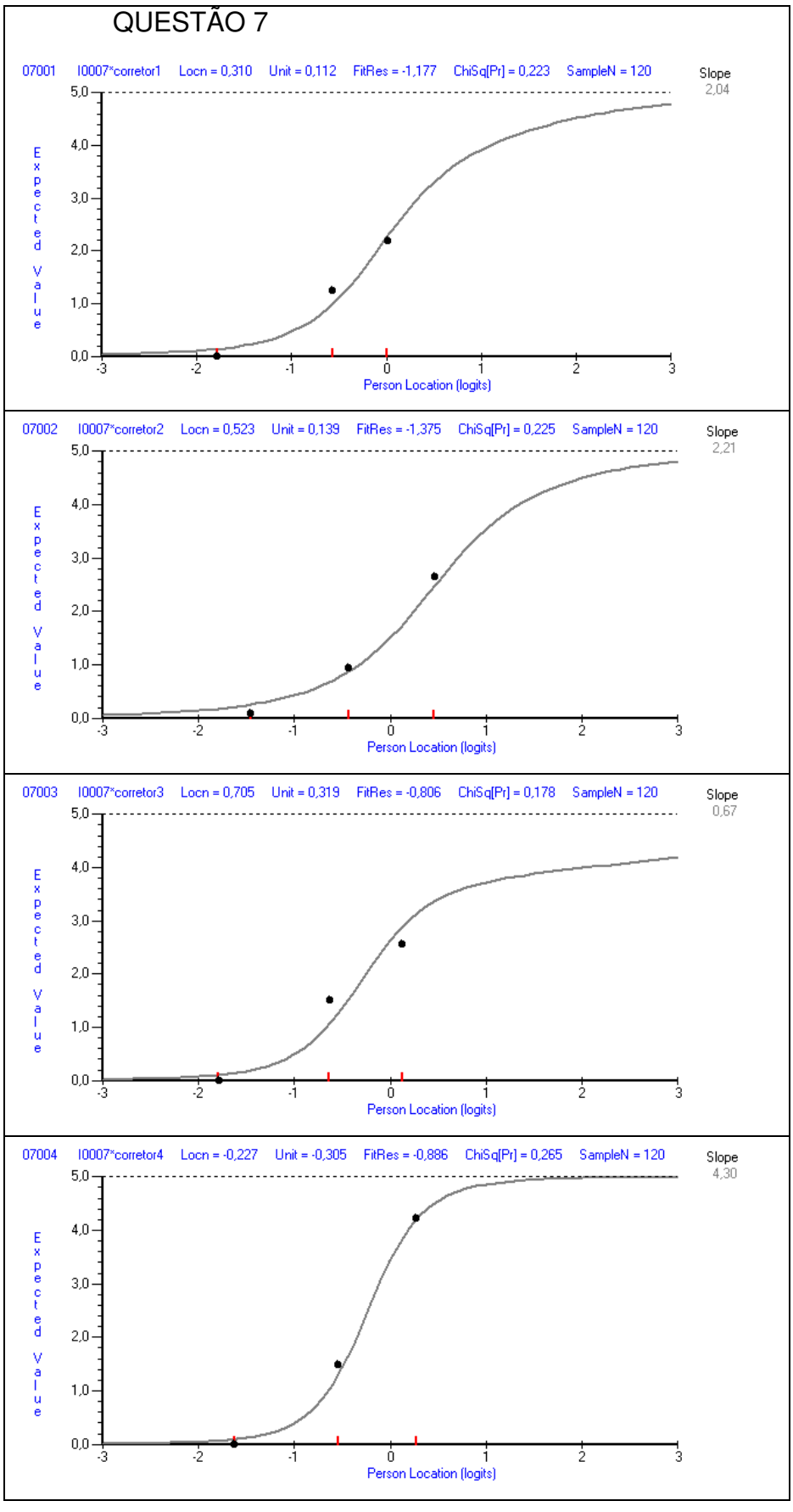
QUESTÃO 5


05001 I0005*corretor1 Locn = -0,415 Unit = 0,317 FitRes = 1,690 ChiSq[Pr] = 0,465 SampleN = 120
Slope 1,54
05002 I0005*corretor2 Locn = 0,120 Unit = 0,319 FitRes = 0,322 ChiSq[Pr] = 0,806 SampleN = 120
Slope 1,03
05003 I0005*corretor3 Locn = 0,166 Unit = 0,363 FitRes = 1,144 ChiSq[Pr] = 0,539 SampleN = 120
Slope 1,19
05004 I0005*corretor4 Locn = -0,769 Unit = 0,160 FitRes = 0,343 ChiSq[Pr] = 0,957 SampleN = 120
Slope 1,05
Expected Value
Person Location (logits)

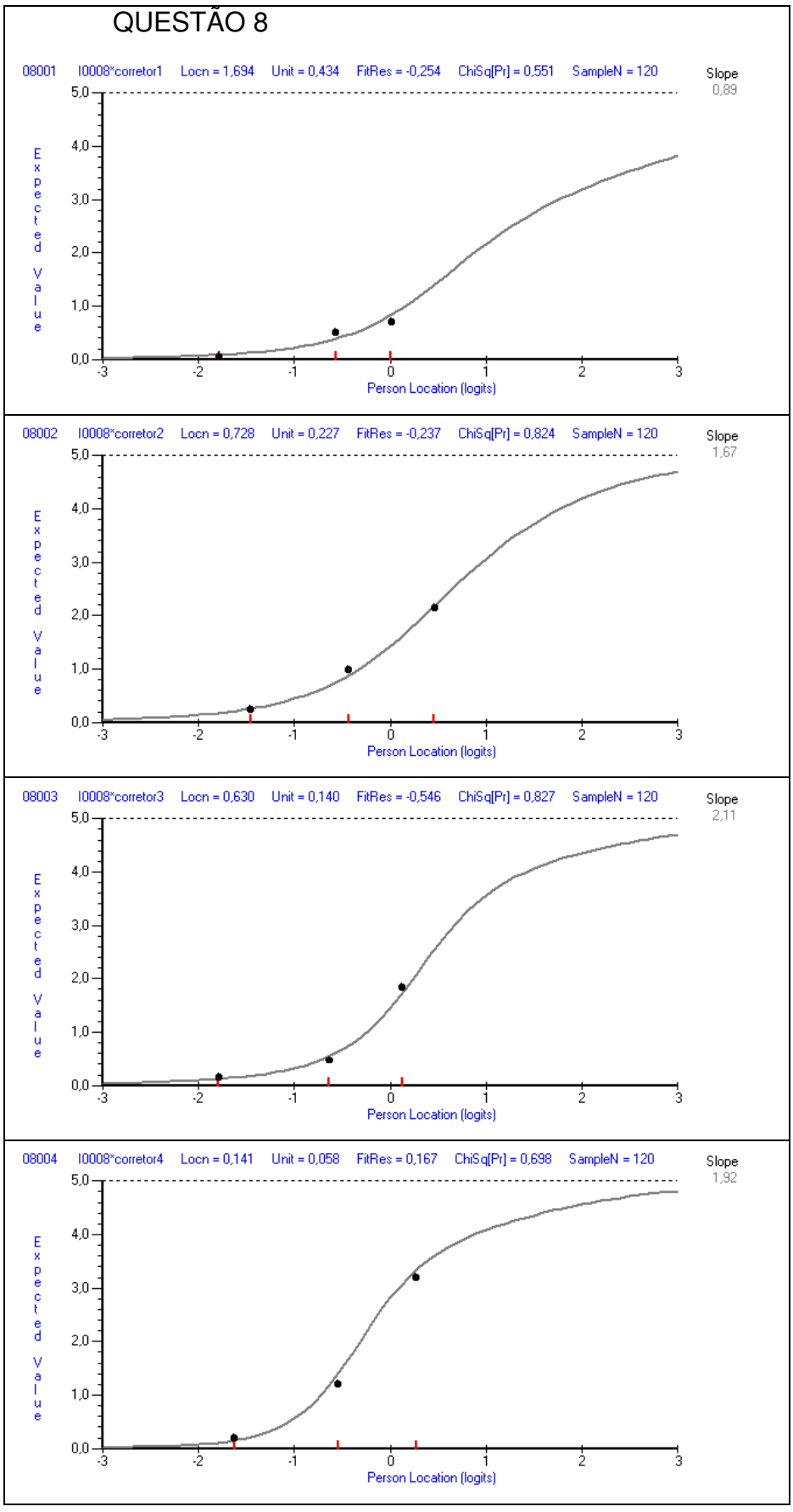
QUESTÃO 6


06001 I0006*corretor1 Locn = 0,342 Unit = 0,365 FitRes = -0,868 ChiSq[Pr] = 0,076 SampleN = 120
Slope
0,98
Expected Value
Person Location (logits)
06002 I0006*corretor2 Locn = 1,231 Unit = 0,563 FitRes = -0,577 ChiSq[Pr] = 0,077 SampleN = 120
Slope
0,90
Expected Value
Person Location (logits)
06003 I0006*corretor3 Locn = -0,233 Unit = 0,301 FitRes = 0,492 ChiSq[Pr] = 0,687 SampleN = 120
Slope
1,34
Expected Value
Person Location (logits)
06004 I0006*corretor4 Locn = 0,675 Unit = 0,526 FitRes = -0,437 ChiSq[Pr] = 0,183 SampleN = 120
Slope
0,82
Expected Value
Person Location (logits)

QUESTÃO 7


07001 I0007*corretor1 Locn = 0,310 Unit = 0,112 FitRes = -1,177 ChiSq[Pr] = 0,223 SampleN = 120
Slope
2,04
07002 I0007*corretor2 Locn = 0,523 Unit = 0,139 FitRes = -1,375 ChiSq[Pr] = 0,225 SampleN = 120
Slope
2,21
07003 I0007*corretor3 Locn = 0,705 Unit = 0,319 FitRes = -0,806 ChiSq[Pr] = 0,178 SampleN = 120
Slope
0,67
07004 I0007*corretor4 Locn = -0,227 Unit = -0,305 FitRes = -0,886 ChiSq[Pr] = 0,265 SampleN = 120
Slope
4,30
Expected Value
5,0
4,0
3,0
2,0
1,0
0,0
-3
-2
-1
0
1
2
3
Person Location (logits)

QUESTÃO 8


08001 I0008*corretor1 Locn = 1,694 Unit = 0,434 FitRes = -0,254 ChiSq[Pr] = 0,551 SampleN = 120
Slope
0,89
Expected Value
Person Location (logits)
08002 I0008*corretor2 Locn = 0,728 Unit = 0,227 FitRes = -0,237 ChiSq[Pr] = 0,824 SampleN = 120
Slope
1,67
Expected Value
Person Location (logits)
08003 I0008*corretor3 Locn = 0,630 Unit = 0,140 FitRes = -0,546 ChiSq[Pr] = 0,827 SampleN = 120
Slope
2,11
Expected Value
Person Location (logits)
08004 I0008*corretor4 Locn = 0,141 Unit = 0,058 FitRes = 0,167 ChiSq[Pr] = 0,698 SampleN = 120
Slope
1,92
Expected Value
Person Location (logits)

ANEXOS

1– REPRODUÇÃO DA TABELA ORIGINAL DOS TIPOS DE PROCESSO COGNITIVO

Table 1.2
Types of Cognitive Processes

---

(1) Recognition. The process of verbatim identification of specific content (e.g., terms, facts, rules, methods, principles, procedures, objects) that was explicitly mentioned in the text.

(2) Recall. The process of actively retrieving from memory and producing content that was explicitly mentioned in the text.

(3) Comprehension. Demonstrating understanding of the text at the mental model level by generating inferences, and interpreting, paraphrasing, translating, explaining, or summarizing information.

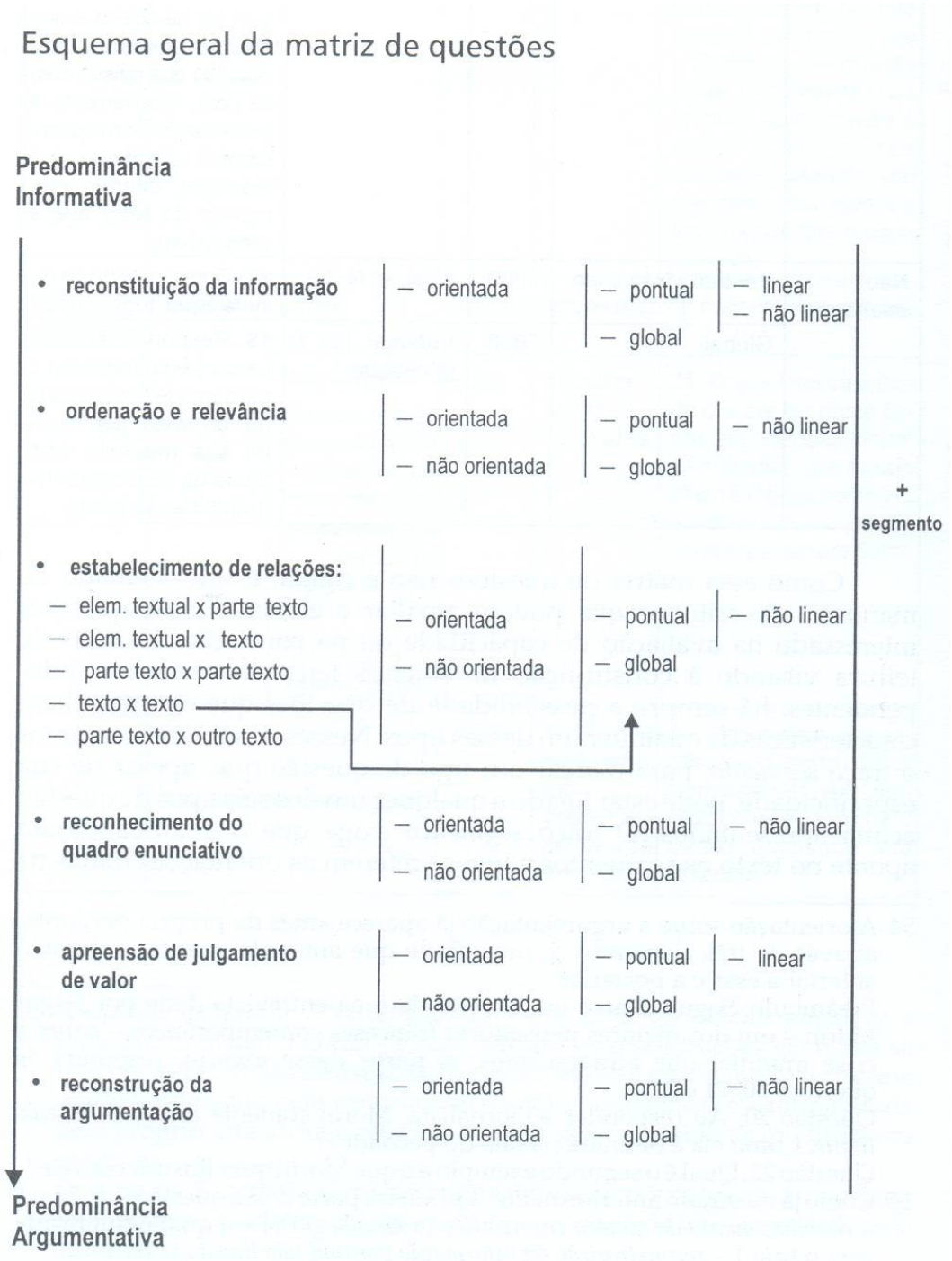
(4) Application. The process of applying knowledge extracted from text to a problem, situation, or case (fictitious or real-world) that was not explicitly mentioned in the text.

(5) Analysis. The process of decomposing elements and linking relationships between elements.

(6) Synthesis. The process of assembling new patterns and structures, such as constructing a novel solution to a problem or composing a novel message to an audience.

(7) Evaluation. The process of judging the value or effectiveness of a process, procedure, or entity, according to some criteria and standards.

2– REPRODUÇÃO DO ESQUEMA GERAL DA MATRIZ DE QUESTÕES



3– PISA - CINCO NÍVEIS DE PROFICIÊNCIA

A escala geral de Leitura representa uma escala síntese dos conhecimentos e habilidades que compõem as três subescalas, distribuídos em cinco níveis de proficiência:

ϕ Nível 1: localizar informações explícitas em um texto, reconhecer o tema principal ou a proposta do autor, relacionar a informação de um texto de uso cotidiano com outras informações conhecidas;

ϕ Nível 2: inferir informações em um texto, reconhecer a idéia principal de um texto, compreender relações, construir sentido e conexões entre o texto e outros conhecimentos da experiência pessoal;

ϕ Nível 3: localizar e reconhecer relações entre informações de um texto, integrar e ordenar várias partes de um texto para identificar a idéia principal, compreender o sentido de uma palavra ou frase e construir relações, comparações, explicações ou avaliações sobre um texto;

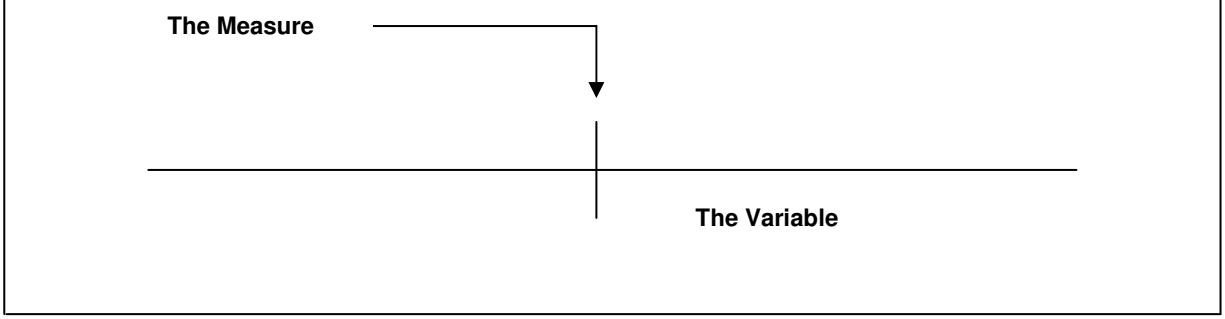
ϕ Nível 4: localizar e organizar informações relacionadas em um texto, interpretar os sentidos da linguagem em uma parte do texto, levando em conta o texto como um todo, utilizar o conhecimento para formular hipóteses ou para avaliar um texto;

ϕ Nível 5: organizar informações contidas, inferindo a informação relevante para o texto, avaliar criticamente um texto, demonstrar uma compreensão global e detalhada de um texto com conteúdo ou forma não familiar.

4– ORIGINAL DA FIGURA 2

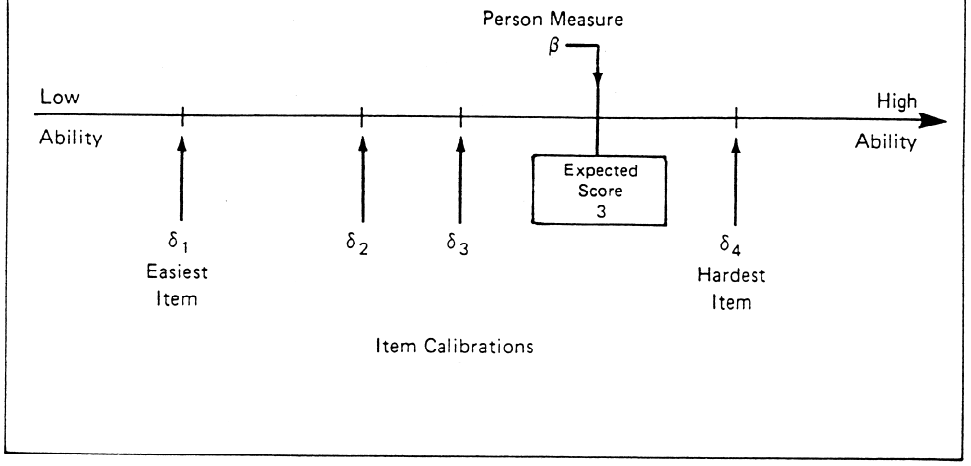
FIGURE 1.1.1

A MEASURE ON A VARIABLE



5– ORIGINAL DA FIGURA 3

FIGURE 1.1.2
DEFINING A VARIABLE



6– EXEMPLOS DE RESUMOS (ABSTRACTS)

(Os resumos estão apresentados segundo a ordem em que aparecem no documento original – texto em português e texto em inglês)

1)

Resumo: [pt]O trabalho agrícola é considerado um dos mais desgastantes visto que exige esforços elevados A operação de máquinas agrícolas requer do trabalhador longas horas na mesma posição executando movimentos repetitivos Este fato leva à fadiga comprometendo a saúde do operador O trabalho do operador de máquinas agrícolas é realizado na cabine a qual muitas vezes apresenta problemas ergonômicos que em conjunto com as condições físicas inapropriadas do operador acabam por comprometer sua saúde e ainda mais a eficiência do trabalho Neste contexto este trabalho foi realizado com o objetivo de realizar o diagnóstico biomecânico, ergonômico e fisiológico de operadores de máquinas agrícolas da Universidade Federal de Lavras e também avaliar se as máquinas utilizadas por eles apresentam as características ergonômicas adequadas Para cada operador foram coletados os dados referentes à pressão arterial VO2máx, índice de massa corporal (IMC), razão cintura/quadril (RCQ) estatura massa percentual de gordura e medidas antropométricas básicas Foram feitos testes de flexibilidade, de preensão manual e o teste de impulsão horizontal Nos tratores foram coletadas as medidas de distância do assento ao volante do assento ao pedal do assento à plataforma e da base do assento Obtiveram-se também os níveis de ruído emitido pelas máquinas Do ponto de vista fisiológico os operadores apresentaram de modo geral bons resultados quanto ao IMC ao percentual de gordura e à pressão arterial No entanto todos apresentaram baixa capacidade física Foi observada também forte correlação entre o IMC e a RCQ Com relação à biomecânica os operadores apresentaram resultados muito fracos para a força de membros inferiores e superiores Quanto à flexibilidade verificou-se um resultado regular considerando a atividade realizada pelos trabalhadores Ergonomicamente os tratores apresentaram valores aceitáveis para os padrões antropométricos dos operadores Com exceção para a distância do assento até a plataforma a qual necessita de correções O nível de pressão sonora das máquinas ficou acima do recomendado para trabalho sem abafador de ruído visto que o tempo de exposição permitido para todas as máquinas é inferior à carga horária de trabalho diário [en]**The agricultural work is considered one of the most stressful since it demands elevated efforts The operation of agricultural machinery formulates a petition of the worker long hours in the same position, executing repetitive movements This fact leads to the fatigue compromising the health of the operator The work of the operator of agricultural machinery is carried out in the cabin being that this one very often presents problems ergonômicos what together with the physical conditions inapropriadas of the operator compromise again still more the efficiency of the work and his health In this**

context the objective of the work carried out the diagnosis biomecânico ergonômico and physiologic preliminary of operators of agricultural machinery of the Federal University of Plowing and also he valued if the machines used by them present the characteristics ergonômicas appropriate For each operator there were collected the data referring to the blood pressure the VO2máx the rate of physical mass (IMC)the reason waist / hip (RCQ) the stature the mass the percentage of fat and the measures antropométricas basic There were done tests of flexibility of manual prehension and the test of horizontal impulse In the tractors there were collected the measures of distance of the seat driving of the seat to the pedal from the seat to the platform and of the base of the seat There were obtained also the levels of noise given out by the machines From the physiologic point of view the operators presented on the whole good results as for the IMC percentage of fat and blood pressure However they all presented low physical capacity Strong correlation was observed also one between the IMC and the RCQ Regarding the biomechanics the operators presented very weak results for the strength of inferior and superior members As for the flexibility a regular result happened considering to activity carried out by the workers Ergonomicamente the tractors presented acceptable values for the standards antropométricos of the operators to the exception for the distance of the seat up to the platform which needs corrections The level of noise of the machines was left above the recommended one for work without the equipment of individual protection (EPI) Being so his use becomes compulsory the time of exhibition allowed for all the machines is inferior to the workload of daily work

http://bibtede.ufla.br/tede//tde_busca/arquivo.php?codArquivo=1299

2)

Resumo: Apresentamos aqui um estudo da Astronomia nova, trabalho publicado pelo astrônomo alemão Johannes Kepler em 1609. O livro é composto na forma de uma narrativa histórica daquela que o astrônomo chamou sua guerra contra Marte, trabalho exaustivo de análise e interpretação dos dados previamente coletados pelo grande observador Tycho Brahe que teve como resultado a descoberta das duas primeiras leis dos movimentos planetários que levam o nome de Kepler. Mostramos que, à luz da Defesa de Tycho contra Ursus, tratado póstumo escrito por Kepler cerca de uma década antes da publicação da Astronomia nova, a estrutura narrativa desta última revela-se como a exposição de um método de pesquisa, segundo o qual o astrônomo percorreu o caminho que leva dos movimentos observados do planeta à determinação de seu percurso real em torno do Sol. Procuramos destacar os principais elementos constituintes deste método, reconstruindo o caminho que leva à descoberta da forma elíptica da órbita do planeta

Abstract: We present an exposition on the New astronomy, published by the german astronomer Johannes Kepler in 1609. The book is composed in the form of a historical narrative of Kepler´s war on Mars, exhaustive work of analysis and interpretation of data relative to the planet previously collected by the great obderver Tycho Brahe, which resulted on the discovery of the two first laws of planetary motion that bear Kepler?s name. We have shown here that in light of Tycho?s defence against Ursus, posthumous work written by Kepler about a decade before the publication of the New astronomy, the historical narrative presented in the latter is the blueprint of a method, by means of which the astronomer derived the true orbit of Mars around the Sun from the observed motions of the planet. We have attempted to provide an account of the main elements of this method, reconstructing the path that leads to the discovery of the elliptical shape of the planet´s orbit

3)

RESUMO O presente trabalho, extraído de uma dissertação do mestrado, explora o viver com o transtorno mental por portadores e familiares, mediante descrição da experiência de oito famílias de portadores de transtorno mental, clientes de hospital-dia, O estudo foi delineado de forma qualitativa com enfoque etnográfico e teve como referencial metodológico os conhecimentos da Antropologia. Ao destacar o significado da experiência, enfatizou-se o discurdo, através do qual foram construídas categorias sob forma de domínios e taxonomia , para deles extrair o tema cultural: Conviver com a doença é difícil! que vem representar o conhecimento cultural dos pesquisados acerca da realidade que os cerca. Os resultados apontam para a necessidade de se repensar a prática em saúde mental nos contextos de família e de se centralizar os cuidados levando em conta os diagnósticos a serem realizados dentro do espaço familiar e comunitário a fim de que se possa construir um modelo de assistência em saúde consoante com a realidade sociocultural das famílias contempladas.

ABSTRACT The presente study looks into the consequences that living with a mental disorder may bring to both patients and their families. The research was done through the intensive insertion of the researcher in the cultural context of the informants, adopting concepts from Anthropology as a methodological reference and giving the data a qualitative treatment under an ethnographic perspective. This study describes and highlights the experiences the families of eight mental disorder patients went through during the period that patients were being treated in the same hospital. The focus is on the research domains and taxonomies. Through these categories the theme Living with a mental disorder is difficult was extracted, which represents the cultural knowledge of the informants about the reality that surroundes them. The results point out to the necessity of re-thinking the practice of mental health services in familiar contexts and of re-focusing these services, taking into account the diagnoses to be made into familiar and community environments, so that a new health care model can be built according to the sociocultural reality of its target families.

http://servicos.capes.gov.br/capesdw/resumo.html?idtese=20027122001018025P5

4)

Resumo: A ciência participa cada vez mais do cotidiano humano ao mesmo tempo em que se torna mais hermética e distante, exercendo importante função social e econômica. As relações entre ciência e sociedade têm sido nas três últimas décadas profundamente marcadas pela multidisciplinaridade, especificidade e aplicabilidade. Esse aumento da esfera de influência da ciência na vida quotidiana tem trazido conseqüências diversas áreas, da geração de energia à saúde, com impacto social e ambiental tanto positivo quanto negativo, o que tem gerado controvérsias e debates. A sociedade tem discutido a relação desenvolvimento científico versus desenvolvimento humano e a ética do fazer científico. Numa sociedade da informação, a imprensa de divulgação científica tem sido palco dessas discussões. Mais que espaço de informação e debate, a atividade de divulgação científica é um canal de educação informal, atingindo um público mais amplo que os bancos escolares. Nesse cenário de inovação e debates a bioética surgiu como uma possível instância de discussão e resposta a esses questionamentos, e a mídia de divulgação científica tem sido importante canal de discussão sobre a ética da ciência e a bioética, inclusive no Brasil. Considerando a importância da atividade de divulgação científica enquanto instância de educação informal, entendemos ser importante estudar a forma como se dá no contexto brasileiro esse processo educativo e de discussão dos temas éticos. No presente trabalho buscamos estudar o discurso sobre ciência, ética e bioética da imprensa brasileira de divulgação científica. Para isso analisamos reportagens das revistas especializadas Galileu e Superinteressante de 2001 tendo como referencial teórico a Análise do Discurso Francesa. Os resultados do estudo empreendido indicam uma tendência prevista por Authier-Revuz: a encenação da participação no processo comunicativo que afasta o leitor de uma participação efetiva no debate científico, silenciando o discurso político e preservando a ciência como instância decisória em questões éticas.

Abstract: Science more and more participates of the human quotidian at the same time it becomes more hermetic and distant, exerting important social and economic function. The relations between science and society have been in the three last decades deeply marked by the multidisciplinarity, especificity and applicability. This increase of the influence sphere of science in the quotidian life have brought consequences in several areas, from the generation of energy to health with social and environmental impact as positive as negative, which has created controversies and discussions. The society has argued the relation scientific development versus human development and the ethics of scientific making. In a society of the information, the press of scientific spreading has been scene of these discussions concerning science and its ethics. More than space of information and debate, the activity of scientific divulgation has been a way of informal education, reaching a public beyond the schools. In this scene of innovation and debates the bioethic took place as a possible instance of discussion and answer to these questionings, and the media of scientific spreading has been an important channel of discussion about the ethics of science and the bioethic, also in Brazil. Considering the importance of the activity of scientific spreading while

instance of informal education, we consider that is important to study the form how this educative process and of discussion of the ethical subjects takes place in the Brazilian context. In this work we search to study the discourse on science, ethics and bioethic of the Brazilian press of scientific spreading. For this we have analyzed news articles of the specialized magazines Galileu and Superinteressante of 2001 having as theorical reference the French Analysis of Discourse. The results of the undertaken study indicate a trend foreseen for Authier- Revuz: the stage of participation on communicative process that moves the reader away from an effective participation on the scientific debate, silencing the political discourse and sustaining the science as a decision instance for questions about ethics

5)

Resumo: A formação do médico é objeto deste estudo, que problematiza os espaços destinados a reflexão sobre os aspectos éticos, que regem o futuro exercício profissional. Repensar o processo cuidar/curar dentro das instituições de ensino e entender como a ética, a moral, os valores fazem parte das vivências escolares e imprimem suas marcas através de uma ou mais disciplinas do currículo foi o eixo da investigação. O método compreendeu entrevistas com docentes, que ministravam a disciplina Temas Longitudinais I, a qual é responsável pela abordagem de conteúdos sobre a ética na dinâmica curricular de um curso de Medicina pertencente a uma Universidade Pública do Estado de São Paulo. Além dos docentes fizeram parte deste estudo alunos, que eram representantes de classe e membros da Comissão de Ensino de Graduação, que estavam envolvidos no processo de reformulação curricular. Os discursos dos sujeitos evidenciam que a dimensão ética deve ser incorporada por todos os responsáveis pelo ensino durante o processo de formação, e não ter uma disciplina apenas como responsável por este enfoque. Assinalam ser este o grande desafio para os docentes do curso de medicina , estar consciente de que o ideal para a formação é ensinar o aluno a aprender a ser e a conviver complementarmente ao aprender a fazer e aprender a aprender. Ressalta-se que os profissionais, que trabalham na área de saúde, mesmo não sendo professores de ética ou de qualquer matéria relacionada, transmitem, uma ética que informa a ação do estudante e se projeta posteriormente no exercício da profissão. Aponta-se ser contraproducente a separação entre o fazer técnico e comportamento ético, pois entende-se, que a competência do médico inclui a competência humana e social. Os docentes devem ser atores ativos neste processo de formação, instigando o educando a uma reflexão constante sobre a ética no seu cotidiano e seu impacto em todo processo formativo. Conclui-se que refletir sobre a dimensão ética na formação do médico requer posicionamento político e homens politicamente posicionados que direcionam os Projetos Políticos Pedagógicos dos cursos de medicina para a aplicação edificante da ciência do cuidar/curar. Um projeto de formação em saúde, deve conter em seus pressupostos esta característica. Deve ser compreendida e assumida coletivamente e tanto quanto possível, ser desenvolvida de forma transdisciplinar. Esta decisão, embora dificultada pelos interesses da lógica de mercado, não pode ser postergada. Trata-se de construir na luta, e a partir das contradições, um projeto contra hegemônico capaz de restaurar a dignidade dos sujeitos, sejam eles, os usuários do sistema de saúde, ou os profissionais que prestam cuidados e que de alguma forma impregnam suas intervenções com valores éticos, os quais são reveladores de uma concepção de Homem verdadeiramente cidadão.

Abstract: The object of this study is the doctor's formation which ethical aspects brings reflection to his future professional career. The basis of this investigation consists in thinking over the healing and caring process in educational institutions and understanding the ethic, the moral and the other values, which take part in the school experiences in one or more disciplines. The method consisted in interviewing the professors who taught Longitudinal Themes I, responsible by the inputs about ethic in

the dynamic of school curriculum belonged to a Medicine course in one of the public universities of São Paulo State. Besides the professors and the university staff, students took part in this study who were class leaders and members of the Undergraduation Commission and were also involved with the curricular reformulation processo. Their speeches show the relevancy of incorporating the ethic in ali the school disciplines during the formation processo They belíeve that this is the biggest challenge to ali the Medicine university staff, since they have to be conscious about the importance of teaching ali the students to do their job and to leam constantly. Even the health professionals who are not ethic masters or any other related subject professor, contribute with their attitudes in the students' future career. It has been considered inefficient to separate the technique and the ethical behavior, since to be a good doctor is necessary to have human and social competence. The teaching staff must be active in the formation process, making the student often think over about ethic in his daily life and its impact during his formation processo. As a conclusion, to reflect about the ethic dimension during the doctor's formation is to make polítical decision in order to manage the Pedagogical Political Projects of Medicine courses to the science use of caring/ healing. A health formation project must follow this characteristic. It must be understood and assumed by everyone, as well as be developed in a trandisciplinary way. This decision, despite the interests of the market logic, must not be postponed. A project against the odds and the hegemony must be built up and it must be able enough to bring back the people's dignity, it does not matter if they are health customers, or health professionals who take care and intervene with their ethical values, which bring out the real conception of Human Being and citizenship.